ECE 8101: Nonconvex Optimization for Machine Learning

Lecture Note 5-1: The Polyak-Łojasiewicz (PL) Condition (feat. Neural Tangent Kernel)

Jia (Kevin) Liu

Assistant Professor
Department of Electrical and Computer Engineering
The Ohio State University, Columbus, OH, USA

Spring 2022

Outline

In this lecture:

- The Polyak-Łojasiewicz (PL) Condition
- Convergence of Various Methods under the PL Condition
- The PL Condition and the Over-parameterized Regime

Convergence Results of Methods We Learned Thus Far

- First-order and zeroth-order methods for nonconvex optimization in learning:
  - GD/SGD-style algorithms
  - Only focus on stationarity gap
  - Typically sublinear convergence rates: $O(1/K)$, $\underbrace{O(1/\sqrt{K})}_{\text{SGD.}}$, ... ($\underbrace{O(1/K^2)}_{\text{Nesterov.}}$ is order-optimal)

    GD (convexity)

    (AGD, ...)

- Meanwhile, it's well-known from convex optimization that:
  - GD achieves linear convergence rate under strong convexity
  - Convergence of global optimality

Can global linear convergence to optimality happen under weaker conditions?

The Polyak-Łojasiewicz Condition

Definition 1 ([Polyak, '63], [Łojasiewicz, '63])

A function $f(\mathbf{x})$ is said to satisfy the Polyak-Łojasiewicz (PL) condition if for all $\mathbf{x} \in \mathbb{R}^d$, there exists a constant $\mu > 0$ such that:

$$2\mu(f(\mathbf{x}) - f(\mathbf{x}^*)) \leq \|\nabla f(\mathbf{x})\|_2^2.$$

Remarks

- Aka "gradient dominated" condition (e.g., [Reddi et al., ICML'16])
- Implies any stationary point is a global min, although not necessarily unique
- Automatically holds for strongly convex functions
- Many nonconvex functions satisfy PL condition, especially in the over-parameterized regime
- PL condition means that the optimality gap $f(\mathbf{x}) - f^*$ is upper bounded by a quadratic function of the stationarity gap

Prop. SC $\to$ PL.

$$f(\underline{y}) \geq f(\underline{x}) + \nabla f(\underline{x})^T (\underline{y} - \underline{x}) + \frac{\mu}{2} \|\underline{y} - \underline{x}\|^2, \ \forall \underline{x}, \underline{y}. \quad \text{(strong convexity)}$$

Minimize both sides w.r.t. $\underline{y}$:

LHS: $f(\underline{x}^*)$

RHS: quad fn of $\underline{y}$: take deriv. w.r.t. $\underline{y}$, set it to 0:

$$\nabla f(\underline{x}) + \mu(\underline{y} - \underline{x}) = 0 \Rightarrow \underline{y}^* = \underline{x} - \frac{1}{\mu} \nabla f(\underline{x}).$$

plugging $\underline{y}^*$ back to RHS:

$$f(\underline{x}) - \frac{1}{\mu} \|\nabla f(\underline{x})\|^2 + \frac{1}{2\mu} \|\nabla f(\underline{x})\|^2 = f(\underline{x}) - \frac{1}{2\mu} \|\nabla f(\underline{x})\|^2$$

So: $f(\underline{x}^*) \geq f(\underline{x}) - \frac{1}{2\mu} \|\nabla f(\underline{x})\|^2$

$$\Rightarrow \|\nabla f(\underline{x})\|^2 \geq 2\mu \left(f(\underline{x}) - f(\underline{x}^*)\right).$$

$\blacksquare$

Nice Features of the PL Condition

- Ease of verification compared to strong convexity (SC):
  - One only needs to access $\|\nabla f(\mathbf{x})\|$ and $f(\mathbf{x})$. In comparison, SC requires checking PD of the Hessian matrix $\mathbf{H}$ (accurate estimation of $\lambda_{\min}(\mathbf{H})$)
- Robustness of the condition
  - $\|\nabla f(\mathbf{x})\|$ is more resilient to perturbation of the obj function than $\lambda_{\min}(\mathbf{H})$
- Allows multiple global minima:
  - Modern ML problems are over-parameterized and have manifolds of global minima, not compatible with SC in general but compatible with PL
- Invariance under transformation:
  - PL is invariant under a broad class of nonlinear coordinate transformations arising from feature extraction/transformation of many ML applications
- PL on manifolds:
  - PL allows for efficient optimization on manifolds, while being invariant under the choice of coordinates (see [Weber and Sra, arXiv:1710:10770]

  Frank-Wolfe
- Linear convergence of GD and SGD:
  - PL is sufficient not only for GD but also for SGD

Gradient Descent under the PL Condition

Theorem 2 (Linear Convergence Rate for GD)

Consider the unconstrained optimization problem $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$*, where* f *has an* L*-Lipschitz continuous gradient, a non-empty solution set* $\mathcal{X}^*$*, and satisfies the PL condition. Then, the gradient descent method with a step-size of* $1/L$*, i.e.,* $\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)$*, has a global linear convergence rate:*

$$f(\mathbf{x}_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(\mathbf{x}_0) - f^*).$$

Remarks

- For twice differentiable functions, L-smoothness means eigenvalues of $\nabla^2 f(\mathbf{x})$ are bounded from above by L (curvature upper bound)

Theorem 2 (Linear Convergence Rate for GD)

Consider the unconstrained optimization problem $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$, *where* f *has an* L*-Lipschitz continuous gradient, a non-empty solution set* $\mathcal{X}^*$, *and satisfies the PL condition. Then, the gradient descent method with a step-size of* $1/L$, *i.e.,* $\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)$, *has a global linear convergence rate:*

$$f(\mathbf{x}_k) - f^* \le \left(1 - \frac{\mu}{L}\right)^k (f(\mathbf{x}_0) - f^*).$$

Proof: GD under PL. $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$.

$$f(x_{k+1}) - f(x_k) \le \nabla f(x_k)^T \underbrace{(x_{k+1} - x_k)}_{-\frac{1}{L}\nabla f(x_k)} + \frac{L}{2}\|\underbrace{x_{k+1} - x_k}_{-\frac{1}{L}\nabla f(x_k)}\|^2.$$

$$\overset{\text{using GD dynamic}}{\le} -\frac{1}{L}\|\nabla f(x_k)\|^2 + \frac{\cancel{L}}{2}\cdot\frac{1}{L^{\cancel{2}}}\|\nabla f(x_k)\|^2$$

$$= -\frac{1}{2L}\|\nabla f(x_k)\|^2$$

$$\overset{PL}{\le} -\frac{1}{\cancel{2}L}\cdot\cancel{2}\mu(f(x_k) - f^*) = -\frac{\mu}{L}(f(x_k) - f^*).$$

Move $f(x_k)$ to RHS, and subtract f^* on both sides:

$$f(x_{k+1}) - f^* \le f(x_k) - f^* - \frac{\mu}{L}(f(x_k) - f^*) = (1 - \frac{\mu}{L})(f(x_k) - f^*). \qquad \blacksquare$$

Exact LS:

$$f(x_{k+1}) = \min_s \{f(x_k) - s\nabla f(x_k)\} \le f(x_k) - \frac{1}{L}\nabla f(x_k).$$

Stochastic Gradient Descent under the PL Condition

- The finite-sum minimization problem: $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{x})$
- Consider the SGD method that uses the iteration: $\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \nabla_{i_k} f(\mathbf{x}_k)$

Theorem 3 (Convergence Rate for SGD)

Assume that f has L-Lipschitz continuous gradients and a non-empty solution set $\mathcal{X}^*$, and it satisfies the PL condition, and f satisfies $\|\nabla f_{i_k}(\mathbf{x}_k)\| \leq C^2$ for all $\mathbf{x}_k$ and some constant $C > 0$. Then, it holds that:

- *SGD with diminishing step-size $s_k = \frac{2k+1}{2\mu(k+1)^2}$ $= O(\frac{1}{\mu k})$. has a convergence rate of:*

$$\mathbb{E}[f(\mathbf{x}_k) - f^*] \leq \frac{LC^2}{2\mu^2 k}. = O(\frac{1}{k}). \quad (\text{as opposed to } O(\frac{1}{\sqrt{k}})).$$

- *SGD with constant step-size $s_k = s \leq \frac{1}{2\mu}$ has a convergence rate of:*

$$\mathbb{E}[f(\mathbf{x}_k) - f^*] \leq (1 - 2s\mu)^k [f(\mathbf{x}_0) - f^*] + \frac{LC^2 s}{4\mu}.$$

Theorem 3 (Convergence Rate for SGD)

Assume that f has L-Lipschitz continuous gradients and a non-empty solution set $\mathcal{X}^*$, and it satisfies the PL condition, and f satisfies $\|\nabla f_{i_k}(\mathbf{x}_k)\| \le C^2$ for all $\mathbf{x}_k$ and some constant $C > 0$. Then, it holds that:

- *SGD with diminishing step-size $s_k = \frac{2k+1}{2\mu(k+1)^2}$ $= O(\frac{1}{\mu k})$. has a convergence rate of:*

$$\mathbb{E}[f(\mathbf{x}_k) - f^*] \le \frac{LC^2}{2\mu^2 k}. = O(\tfrac{1}{k}). \quad \left(\text{as opposed to } O(\tfrac{1}{\sqrt{k}})\right).$$

- *SGD with constant step-size $s_k = s \le \frac{1}{2\mu}$ has a convergence rate of:*

$$\mathbb{E}[f(\mathbf{x}_k) - f^*] \le (1 - 2s\mu)^k [f(\mathbf{x}_0) - f^*] + \frac{LC^2 s}{4\mu}.$$

Proof. $f(x_{k+1}) \le f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2} \|\nabla f_{i_k}(x_k)\|^2.$

$\overset{\text{SGD dynamic}}{=} f(x_k) + s_k \nabla f(x_k)^T \nabla f_{i_k}(x_k) + \frac{L s_k^2}{2} \|\nabla f_{i_k}(x_k)\|^2.$

Take full expectation w.r.t. $\{i_k\}$:

$$\mathbb{E}[f(x_{k+1})] \le \mathbb{E}[f(x_k)] - s_k \mathbb{E}[\nabla f(x_k)^T \nabla f_{i_k}(x_k)] + \frac{L s_k^2}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

$$\le \mathbb{E}[f(x_k)] - 2 s_k \mathbb{E}[\|\nabla f(x_k)\|^2] + \frac{L C^2 s_k^2}{2}$$

$$\overset{PL}{\le} \mathbb{E}[f(x_k)] - 2\mu s_k \mathbb{E}[f(x_k) - f^*] + \frac{L C^2 s_k^2}{2}.$$

Subtracting f^* on both sides:

$$\mathbb{E}[f(x_{k+1}) - f^*] \le (1 - 2\mu s_k) \mathbb{E}[f(x_k) - f^*] + \frac{L C^2 s_k^2}{2}. \quad (1).$$

1° diminishing step-size: $s_k = \frac{2k+1}{2\mu(k+1)^2} = O(\frac{1}{\mu k}).$

$$\mathbb{E}[f(x_{k+1}) - f^*] \le \frac{k^2}{(k+1)^2} \mathbb{E}[f(x_k) - f^*] + \frac{L C^2 (2k+1)^2}{8\mu^2 (k+1)^4}$$

Multiplying both sides by $(k+1)^2$ and letting $\delta_f(k) = k^2 \mathbb{E}[f(x_k) - f^*]$.

$$\delta_f(k+1) \le \delta_f(k) + \frac{Lc^2 (2k+1)^2}{8\mu^2 (k+1)^4} \le \delta_f(k) + \frac{Lc^2}{2\mu^2} \quad (2)$$

$$\underbrace{\frac{2k+1}{k+1}}_{<2}$$

Summing (2) from $k=0$ to $K-1$ and $\delta_f(0) = 0^2 \mathbb{E}[f(x_0) - f^*] = 0$.

$$\delta_f(k) \le \underbrace{\delta_f(0)}_{=0} + \frac{Lc^2 k}{2\mu^2} \Rightarrow k^2 \mathbb{E}[f(x_k) - f^*] \le \frac{Lc^2 \cancel{k}}{2\mu^2}.$$

$$\Rightarrow \mathbb{E}[f(x_k) - f^*] \le \frac{Lc^2}{2k\mu^2}.$$

2° Constant step-size: $s_k = s$ for some $s < \frac{1}{2\mu}$. Applying in (1) recursively.

$$\mathbb{E}[f(x_{k+1}) - f^*] \le (1-2\mu s)^k \mathbb{E}[f(x_0) - f^*] + \frac{Lc^2 s^2}{2} \sum_{i=0}^{k} (1-2s\mu)^i$$

$$\le (1-2\mu s)^k \mathbb{E}[f(x_0) - f^*] + \frac{Lc^2 s^2}{2} \sum_{i=0}^{\infty} (1-2s\mu)^i$$

$$= (1-2\mu s)^k \mathbb{E}[f(x_0) - f^*] + \underbrace{\frac{Lc^2 s^2}{2} \cdot \frac{1}{2\mu s}}_{= \frac{Lc^2 s}{4\mu}}$$

∎

SGD under PL Condition in Over-parameterized Regime

- Consider ERM in over-parameterized regime: $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{x})$
  - $f(\mathbf{x})$ is L-smooth: $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|$, $\forall \mathbf{x},\mathbf{y}$
  - $f_i(\mathbf{x})$ satisfies: $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq \tilde{L}|f_i(\mathbf{x}) - f_i(\mathbf{y})|$ for some $\tilde{L} > 0$
  - In ML problems, w.l.o.g., we can assume that $\inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) = 0$ and so the PL condition can be modified as μ-PL*: $2\mu f(\mathbf{x}) \leq \|\nabla f(\mathbf{x})\|_2^2$

Implied by Lipschitz-cont. and smoothness of fi.

$\|\nabla f(\mathbf{x})\|^2 \geq 2\mu(f(\mathbf{x}) - \underbrace{f^*}_{=0})$

- Over-parameterized regime: $d \gg N$
  - The interpolation effect: for every sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots$ such that $\lim_{k\to\infty} f(\mathbf{x}_k) = 0$, we have

$$\lim_{k\to\infty} f_i(\mathbf{x}_k) = 0, \quad 1 \leq i \leq N.$$

  - Meaning: In the over-parameterized regime, the richness of the model is so high such that fit all training samples

$A\mathbf{x} = \mathbf{b}$, $A \in \mathbb{R}^{N\times d}$, $\mathbf{x}\in\mathbb{R}^d$ $\quad f = \min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2$.

$N > d$. Null space : $N-d$ dimension.

SGD under PL Condition in Over-parameterized Regime

- Consider the general mini-batched version of SGD with constant step-size s:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{s}{B} \sum_{j=1}^{B} \nabla f_{i_k^j}(\mathbf{x}_k),$$

  - B: the mini-batch size; the sample indices $\{i_k^1, \ldots, i_k^B\}$ in the mini-batch are drawn uniformly with replacement in each iteration k from $\{1, \ldots, N\}$

Theorem 4 ([Bassily et al., arXiv:1811.02564])

Consider the mini-batch SGD with smooth losses as stated. Suppose the interpolation condition holds. Suppose that the ERM function $f(\mathbf{x})$ is μ-PL for some $\mu > 0$. For any mini-batch size $B \in \mathbb{N}$, the mini-batch SGD with constant step-size $s^*(B) \triangleq \frac{2\mu B}{L(\tilde{L}+L(B-1))}$ guarantees that:*

$$\mathbb{E}[f(\mathbf{x}_k)] \leq (1 - \mu s^*(B))^k f(\mathbf{x}_0)$$

Theorem 4 ([Bassily et al., arXiv:1811.02564])

Consider the mini-batch SGD with smooth losses as stated. Suppose the interpolation condition holds. Suppose that the ERM function $f(\mathbf{x})$ is μ-PL for some $\mu > 0$. For any mini-batch size $B \in \mathbb{N}$, the mini-batch SGD with constant step-size $s^*(B) \triangleq \frac{2\mu B}{L(\tilde{L}+L(B-1))}$ guarantees that:*

$$\mathbb{E}[f(\mathbf{x}_k)] \leq (1 - \mu s^*(B))^k f(\mathbf{x}_0)$$

Proof: From descent lemma:

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2.$$

Using SGD dynamics:

$$f(x_{k+1}) - f(x_k) \leq -s \nabla f(x_k)^T \left[\frac{1}{B}\sum_{j=1}^{B} \nabla f_{i_k^j}(x_k)\right] + \frac{Ls^2}{2} \left\|\frac{1}{B}\sum_{j=1}^{B} \nabla f_{i_k^j}(x_k)\right\|^2.$$

Fix x_k, and take expectation over choice $\{i_k^1, \cdots, i_k^B\}$.

Note: indices are i.i.d., Then, we have:

$$\mathbb{E}\left[f(x_{k+1}) - f(x_k) \,\middle|\, x_k\right] \leq -s \|\nabla f(x_k)\|^2 + \frac{s^2 L}{2}\left(\frac{1}{B} \mathbb{E}_{i_k^1}\left[\|\nabla f_{i_k^1}(x_k)\|^2\right] + \frac{B-1}{B} \|\nabla f(x_k)\|^2\right).$$

Since $\forall i \in \{1, \cdots, N\}$, $f_i(\cdot)$ satisfies:

$$\|\nabla f_{i_k^1}(x_k) - \underbrace{\nabla f_{i_k^1}(x^*)}_{=0}\| \leq \tilde{L} \|f_{i_k^1}(x_k) - \underbrace{f_{i_k^1}(x^*)}_{=0}\| = \tilde{L} \|\underbrace{f_{i_k^1}(x_k)}_{\geq 0}\| \leq 2\tilde{L} f_{i_k^1}(x_k)$$

Thus.

$$\mathbb{E}\left[f(x_{k+1}) - f(x_k) \,\middle|\, x_k\right] \leq -s\left(1 - \frac{sL}{2} \cdot \frac{B-1}{B}\right) \|\nabla f(x_k)\|^2 + \frac{s^2 L \tilde{L}}{B} f(x_k).$$

$$\underset{\substack{\text{non-neg.}\\ f(x)}}{\overset{PL}{\leq}} 2\mu s\left(1-\frac{sL(B-1)}{B}\right)f(x_k)+\frac{s^2L\tilde{L}}{B}f(x_k)$$

$$= s\left[-2\mu-\frac{sL}{B}\left(\mu(B-1)+\tilde{L}\right]f(x_k).\right.$$

Finally, rearranging & taking full expectation:

$$\mathbb{E}\left[f(x_{k+1})\right] \leq \left[1-2s\mu+\frac{s^2L}{B}(\mu(B-1)+\tilde{L}\right]\mathbb{E}\left[f(x_k)\right]. \quad (3).$$

Optimizing the quad term in (3) w.r.t. s yield $s^*(B)$,

$$\mathbb{E}\left[f(x_{k+1})\right] \leq (1-\mu s^*(B))\,\mathbb{E}\left[f(x_k)\right]. \qquad \blacksquare$$

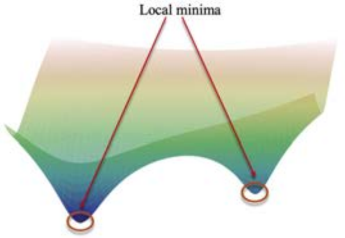
Other Methods under the PL Condition

Similar linear convergence rate results can be shown for other methods under the μ-PL, L-smoothness, and uniform variance bound conditions, which implies the following sample complexity results:

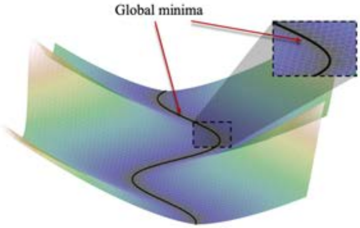
- GD [Polyak, '63]: $\frac{L}{\mu}\log\frac{\Delta_0}{\epsilon}$
- SGD [Karimi et al., ECML-KDD'16]: $\frac{L}{\mu}(\frac{\max_i L_i}{\mu}\log(\frac{\Delta_0}{\epsilon}) + \frac{\max_i L_i \Delta_*}{\mu\epsilon})$
- SVRG [Reddi et al., NeurIPS'16]: $(N + \frac{N^{2/3}\max_i L_i}{\mu})\log(\frac{\Delta_0}{\epsilon})$
- SAGA [Reddi et al., NeurIPS'16]: $(N + \frac{N^{2/3}\max_i L_i}{\mu})\log(\frac{\Delta_0}{\epsilon})$
- PAGE [Li et al., ICML'21]: $(b + \sqrt{b}\frac{L_{\text{avg}}}{\mu})\log(\frac{\Delta_0}{\epsilon})$, where $b = \min\{\frac{\sigma^2}{\mu\epsilon}, N\}$

PL Condition and Over-parameterized Regime

- Landscape of under-parameterized and over-parameterized models (figure from [Liu et al., arXiv:2003:00307]



(a) Loss landscape of under-parameterized models



(b) Loss landscape of over-parameterized models

$d \gg N.$

- Key Insight:
  - Convexity is not the right framework for analyzing the loss landscape of over-parameterized systems, even locally
  - Instead, the μ-PL* condition (i.e., $\|\nabla f(\mathbf{w})\|_2^2 \geq 2\mu f(\mathbf{w})$, $\forall \mathbf{w}$) is a more appropriate framework

PL Condition and Over-parameterized Regime

The essence of supervised learning:

- Given a dataset of size N, $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, $\mathbf{x}_i \in \mathbb{R}^d$, $y \in \mathbb{R}$
- A parametric family of models $f(\mathbf{w}, \mathbf{x})$ (e.g., a neural network)
- Goal: To find a model with parameter $\mathbf{w}^*$ that fits the training data:

$$f(\mathbf{w}^*, \mathbf{x}_i) \approx y_i, \quad i = 1, 2, \ldots, N$$

- Mathematically: Equivalent to solving (exactly or approximately) a system of N nonlinear equations:

vector - valued

$$\mathcal{F}(\mathbf{w}) = \mathbf{y},$$

where $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{y} \in \mathbb{R}^N$, and $\mathcal{F}(\cdot) : \mathbb{R}^d \to \mathbb{R}^N$ with $(\mathcal{F}(\mathbf{w}))_i = f(\mathbf{w}, \mathbf{x}_i)$.

- The system of equations is solved by minimizing a certain loss function $\mathcal{L}(\mathbf{w})$
  - E.g., the square loss: $\mathcal{L}(\mathbf{w}) = \frac{1}{2}\|\mathcal{F}(\mathbf{w}) - \mathbf{y}\|^2 = \frac{1}{2}\sum_{i=1}^{N}(f(\mathbf{w}, \mathbf{x}_i) - y_i)^2$

PL Condition and Over-parameterized Regime

μ-PL* condition emerges through the spectrum of the tangent kernel

- Let $D\mathcal{F}(\mathbf{w}) \in \mathbb{R}^{N\times d}$ be the differential of the mapping $\mathcal{F}$ at $\mathbf{w}$
- The tangent kernel of $\mathcal{F}$ is defined as an $N \times N$ matrix:

$$\mathbf{K}(\mathbf{w}) \triangleq D\mathcal{F}(\mathbf{w})D\mathcal{F}^\top(\mathbf{w})$$

  - It follows from the definition that $\mathbf{K}(\mathbf{w})$ is PSD

$\|\nabla f(x)\| \geq 2\mu(f(x))$.

- The square loss $\mathcal{L}$ is μ-PL* at $\mathbf{w}$ [Liu, et al., arXiv:2003:00307], where

$\|\nabla f(x) - \underbrace{\cancel{\nabla f(x^*)}}_{=0}\| \geq 2\mu(f(x) - \underbrace{\cancel{f^*}}_{=0})$

$$\mu = \lambda_{\min}(\mathbf{K}(\mathbf{w})),$$

is the smallest eigenvalue of the kernel matrix

Thus, the PL* condition is inherently tied to the spectrum of the tangent kernel matrix associated with $\mathcal{F}$

PL Condition and Over-parameterized Regime

Wide (hence over-parameterized) neural networks satisfy PL* condition:

- A powerful tool: the neural tangent kernel (NTK)
  - First appeared in a landmark paper [Jacot et al., NeurIPS'18]
  - Tangent kernel of a single-hidden layer wide neural networks with linear output layer ($f(\mathbf{x}) = \sum_{i=1}^{d} \sigma(\mathbf{w}^\top \mathbf{x})$) are nearly constant in a ball $\mathcal{B}$ of a certain radius around the ball with a random center (note: d is also the width of the NN):

$$\|\mathbf{H}_{\mathcal{F}}(\mathbf{w})\| = O^*(1/\sqrt{d}),$$

    where $\mathbf{H}_{\mathcal{F}}(\mathbf{w})$ is a $N \times d \times d$ tensor with $(\mathbf{H}_{\mathcal{F}})_{ijk} = \frac{\partial^2 \mathcal{F}_i}{\partial w_j \partial w_k}$
  - Constancy of NTK implies training dynamic of wide NNs is approximately a linear model $\Rightarrow$ linear convergence of gradient flow (hence GD)
  - It can be shown that [Liu, et al., arXiv:2003:00307]:

$$|\lambda_{\min}(\mathbf{K}(\mathbf{w})) - \lambda_{\min}(\mathbf{K}(\mathbf{w}_0))| < O\left(\sup_{\mathbf{w} \in \mathcal{B}} \|\mathbf{H}_{\mathcal{F}}(\mathbf{w})\|\right) = O(1/\sqrt{d})$$

Thus, the PL* condition holds for single-layer wide NN

NTK:

High-level intuition:

1° If loss is convex, then GD (SGD) converges to global min.

2° Linear/kernel model exhibit convex loss landscape.

3° Will prove wide NN landscape looks like a kernel model.

$$3^\circ \to 2^\circ \to 1^\circ$$

1) Gradient dynamic for linear models:

Dataset $\{(\underline{x}_i, y_i)\}_{i=1}^N$ $\quad u_i = \underline{w}^T \underline{x}_i \quad \underline{x}_i \in \mathbb{R}^d$

$$\underline{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_N \end{bmatrix} \in \mathbb{R}^N \quad \underline{\underline{X}} = \begin{bmatrix} \text{---} & \underline{x}_1^T & \text{---} \\ & \vdots & \\ \text{---} & \underline{x}_N^T & \text{---} \end{bmatrix}_{N\times d} \quad \underline{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} \quad d \gg N$$

Square loss: $L(\underline{w}) = \frac{1}{2}\sum_{i=1}^N (y_i - \underline{x}_i^T \underline{w})^2 = \frac{1}{2}\|\underline{y} - \underline{\underline{X}}\,\underline{w}\|^2$

Optimize via GD: $\quad \underline{w}^+ = \underline{w} - s\nabla L(\underline{w})$

$$\nabla_{\underline{w}} L(\underline{w}) = -\sum_{i=1}^N \underline{x}_i (y_i - u_i) = -\underline{\underline{X}}^T(\underline{y} - \underline{u})$$

$s \to 0$: $\quad \dfrac{d\underline{w}}{dt} = -\nabla L(\underline{w}) = \underline{\underline{X}}^T(\underline{y} - \underline{u}) \quad$ (ODE for $\underline{w}$ evolution)

Gradient flow.

GD: a finite-time discretization of this ODE

$$\frac{d\underline{u}}{dt} = \frac{d(\underline{\underline{X}}\,\underline{w})}{dt} = \underline{\underline{X}}\frac{d\underline{w}}{dt} = \underbrace{\underline{\underline{X}}\,\underline{\underline{X}}^T}_{\underline{\underline{K}}}(\underline{y} - \underline{u}) = \underline{\underline{K}}(\underline{y} - \underline{u}).$$

Remarks:

1° Linear ODE can be solved in closed form. Let $\underline{r} = y - \underline{u}$

$$\frac{d\underline{r}}{dt} = \frac{d(y-\underline{u})}{dt} = -\frac{d\underline{u}}{dt} = -\underline{\underline{K}}\,\underline{r}$$

$$\underline{r}(t) = \exp(-\underline{\underline{K}}\,t)\,\underline{r}(0)$$

If $\underline{\underline{K}}$ is full rank, $\lambda_{min}(\underline{\underline{K}}) > 0$.

GD converges exp. fast $\to$ 0 loss.

2° $\underline{\underline{K}} = \underline{\underline{X}}\,\underline{\underline{X}}^T$ const. Configuration of data pts with $\lambda_{min}(\underline{\underline{K}})$ allows GD to converge fast.

3° All that matters is set of pair wise product $[\underline{\underline{K}}]_{ij} = \underline{x}_i^T \underline{x}_j$

"kernel trick". Kernel fn:

$[\underline{\underline{K}}]_{ij} = \langle \phi(\underline{x}_i), \phi(\underline{x}_j) \rangle$, where ϕ is "feature map".

2). General dynamics for non-linear model: $f(\underline{w})$.

For $\underline{x}_i$, $u_i = f(\underline{w}, \underline{x}_i)$.

Square loss: $L(\underline{w}) = \frac{1}{2}\sum_{j=1}^{N}(y_j - \underbrace{f(\underline{w}, \underline{x}_j)}_{u_j})$

The grad w.r.t. any one weight parameter:

$$\nabla_{w_i} L(\underline{w}) = -\sum_{j=1}^{N} \frac{\partial f(\underline{w}, \underline{x}_i)}{\partial w_i}(y_j - u_i).$$

$$\frac{du_i}{dt} = \sum_{k=1}^{d} \frac{\partial u_i}{\partial w_k}\frac{dw_k}{dt} = \sum_{k=1}^{d} \frac{\partial f(\underline{w}, \underline{x}_i)}{\partial w_k}\frac{dw_k}{dt}$$

$$= \sum_{k=1}^{d} \frac{\partial f(\underline{w}, \underline{x}_i)}{\partial w_k} \left[\sum_{j=1}^{N} \frac{\partial f(\underline{w}, \underline{x}_j)}{\partial w_k} (y_j - u_j) \right]$$

$$= \sum_{j=1}^{N} \left\langle \frac{\partial f(\underline{w}, \underline{x}_i)}{\partial \underline{w}}, \frac{\partial f(\underline{w}, \underline{x}_j)}{\partial \underline{w}} \right\rangle (y_j - u_j)$$

$$= \sum_{j=1}^{N} [\underline{\underline{K}}]_{ij} (y_j - u_j), \quad \text{where } [\underline{\underline{K}}]_{ij} = \left\langle \frac{\partial f(\underline{w}, \underline{x}_i)}{\partial \underline{w}}, \frac{\partial f(\underline{w}, \underline{x}_j)}{\partial \underline{w}} \right\rangle$$

$$= \sum_{k=1}^{d} \frac{\partial f(\underline{w}, \underline{x}_i)}{\partial w_k} \cdot \frac{\partial f(\underline{w}, \underline{x}_j)}{\partial w_k}.$$

$$\underline{\underline{K}}_t = DF(\underline{w}_t) DF^T(\underline{w}_t).$$ kernel matrix.

$$\frac{d\underline{u}}{dt} = -\underline{\underline{K}}_t (\underline{y} - \underline{u}).$$ ← nonlinear ODE.

1°. $\underline{\underline{K}}_t \succeq 0, \ \forall t.$ kernel mapping: $\phi: \underline{x} \longmapsto \frac{\partial f(\underline{w}, \underline{x})}{\partial \underline{w}} \in \mathbb{R}^d$

If f is NN, then ϕ is "NTK".

3). Wide NN exhibits linear model dynamics. [Du, ICLR'19]

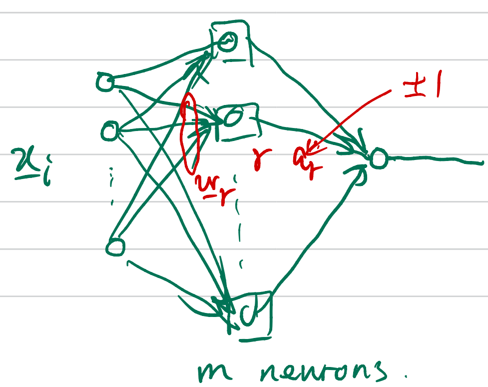
1° Randomly initialize $\underline{w}$ at $t=0$.

2° At $t=0$, we'll show NTK $\underline{\underline{K}}_0$ is full rank.

3° For wide NNs, $\underline{\underline{K}}_t \approx \underline{\underline{K}}_0$, hence $\underline{\underline{K}}_t$ is full rank $\forall t$.

Consider 2-layer NN w/ m hidden neurons, with twice diff'ble ψ activation fn.

Fix 2nd layer, only train 1st layer.



$$f(\underline{w}, \underline{x}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \psi(\langle \underline{w}_r, \underline{x} \rangle) \quad , \quad a_r = \pm 1$$

Initialize $[w_1(0) \cdots w_m(0)]^T$ stand normal distr.

$$\frac{\partial f(\underline{w}(0), \underline{x}_i)}{\partial \underline{w}_r} = \frac{1}{\sqrt{m}} a_r \underline{x}_i \psi'(\langle \underline{w}_r(0), \underline{x}_i \rangle).$$

So, NTK at $t=0$: $[\underline{\underline{K}}]_{ij} = \left\langle \frac{\partial f(\underline{w}, \underline{x}_i)}{\partial \underline{w}_r}, \frac{\partial f(\underline{w}, \underline{x}_j)}{\partial \underline{w}_r} \right\rangle$.

$$= \underline{x}_i^T \underline{x}_j \left[\frac{1}{m} \sum_{r=1}^{m} a_r^2 \psi'(\langle \underline{w}_r(0), \underline{x}_i \rangle) \psi'(\langle \underline{w}_r(0), \underline{x}_j \rangle) \right]$$

Each entry of $[\underline{\underline{K}}]_{ij}$ is a r.v. with mean being equal to:

$$\underline{x}_i^T \underline{x}_j \, \mathbb{E}_{\underline{w} \sim N(0, \underline{\underline{I}})} \psi'(\underline{x}_i^T \underline{w}) \psi'(\underline{x}_j^T \underline{w}) \triangleq [\underline{\underline{K}}^*]_{ij}$$

As $m \to \infty$, NTK at $t=0$ is qual to $\underline{\underline{K}}^*$

1°. for $\varepsilon > 0$ if $m > \tilde{O}\left(\frac{N^4}{\varepsilon^2}\right)$. then $\|\underline{\underline{K}}(0) - \underline{\underline{K}}^*\| \le \varepsilon$ w.h.p.

2° Suppose $y_i = \pm 1$, $u_i(\tau)$ bnded throughout training, $0 \le \tau \le t$.

for $\varepsilon > 0$. if $m \ge \tilde{O}\left(\frac{N^6 t^2}{\varepsilon^2}\right)$, then $\|\underline{\underline{K}}(\tau) - \underline{\underline{K}}^*\| \le \varepsilon$, w.h.p.

Remarks:

(1). width scales ploy w.r.t. N. [Song et al. NeurIPS'21] $\tilde{O}(N^{\frac{3}{2}})$

(2). Dependence on data: [Nguyen et al.]: $O(Nd)$

(3) For L-layer NN, widths need to scale as poly(N, L).

Next Class

First-Order Methods under Additional Assumptions