SGD: A Stability Perpective

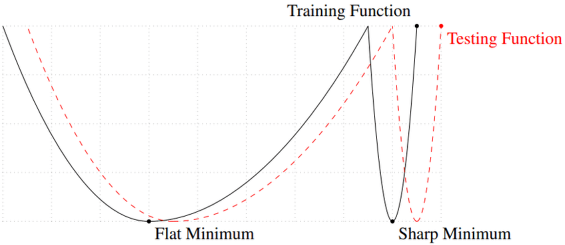
Arjun Ashok Rao

December 13, 2021

- Large Neural Networks (NNs) trained with SGD easily interpolate the training data.
- [Zhang et al., 2016]: In the overparameterized regime, NNs easily fit random labels with zero training error.
- Deep NN models also have $\gg 1$ Global minima.
- (New) Role of optimization: Among all the the global minima with zero training error, which global minima produces zero *test* error.

- Large Neural Networks (NNs) trained with SGD easily interpolate the training data.
- [Zhang et al., 2016]: In the overparameterized regime, NNs easily fit random labels with zero training error.
- Deep NN models also have $\gg 1$ Global minima.
- (New) Role of optimization: Among all the the global minima with zero training error, which global minima produces zero *test* error.
- New Question: What are the characteristics of (global) minima that (a) SGD can converge to, and (b) generalize well?

- New Question: What are the characteristics of (global) minima that (a) SGD can converge to, and (b) generalize well?
- Traditional Answer: Flat Minima, since loss landscapes of train and test data is similar **upto** a certain perturbation.



- Incorrect , [Dinh et al., 2017] carefully re-parameterize and disprove this with Volume, Hessian-based flatness measures
- New Answer: SGD is biased to flat-minima solutions.

From Flatness to Stability: Toy Example with Quadratics

- Consider the one-dimensional quadratic $f(x) = \frac{1}{2}ax^2 + bx + c, \quad a > 0$, optimum given by $x^* = \frac{-b}{a}$
- Update rule with vanilla Gradient Descent:

$$\begin{aligned} x_{t+1} &= x_t - \eta \nabla f(x) && (1) \\ &= x_t - \eta(ax_t + b) && (2) \\ \Rightarrow x_{t+1} - x^* &= (1 - \eta a)(x_t - x^*) && (3) \\ \therefore x_t &= (1 - \eta a)^t (x_0 - x^*) + x^* && (4) \end{aligned}$$

- If $a \geq \frac{2}{\eta}$, $(1 - \eta a) < -1$, divergence.

- A Generalization: Consider $f(x) = \frac{1}{2}x^T \mathsf{A} x + \mathsf{b}^T x + c$. Let (q, a) be an eigenvalue, eigenvector pair of A.

$$x_{t+1} = x_t - \eta(\mathsf{A}x_t + \mathsf{b}) = (\mathbb{I} - \eta\mathsf{A})x_t - \eta\mathsf{b} \tag{5}$$

- Consider the quantity $\mathsf{q}^T x_t$

$$\mathsf{q}^T x_{t+1} = \mathsf{q}^T(\mathbb{I} - \eta\mathsf{A})x_t - \eta\mathsf{b} \tag{6}$$

$$= (1 - \eta a)\mathsf{q}^T x_t - \eta\mathsf{q}^T\mathsf{b} \qquad (\mathsf{q}^T\mathsf{A} = a\mathsf{q}) \tag{7}$$

- Since $\eta > 0$, if $a \geq \frac{2}{\eta}$, then $(1 - \eta a) < -1$, $\mathsf{q}^T x_t$ will diverge.
- Intuition: For NN, 2^{nd} order Taylor approximation near initialization point $\boldsymbol{\theta}_0$ is a quadratic function. Note here that A in this case will be equivalent to the Hessian.

- Consider a loss function parameterized by $\boldsymbol{\theta}$ for stochastically sampled data given by:

$$\hat{L}_t(\boldsymbol{\theta}) = \frac{1}{B} \sum_{j \in \mathcal{B}_t} l_j(\boldsymbol{\theta})$$

- $l : \mathbb{R}^d \to \mathbb{R}$ is differentiable $\forall j \in [n]$. Consider a twice-differentiable minima $\boldsymbol{\theta}^*$.

$$\hat{L}_t(\boldsymbol{\theta}) \approx \hat{L}_t(\boldsymbol{\theta}^*) + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \nabla \hat{L}_t(\boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \nabla^2 \hat{L}_t(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \quad (8)$$

Definition — Linear Stability [Mulayoff et al., 2021]

If $\boldsymbol{\theta}^*$ is a twice differentiable minima of L, and the following linearized stochastic dynamical system applies:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta(\nabla \hat{L}_t(\boldsymbol{\theta}^*) + \nabla^2 \hat{L}_t(\boldsymbol{\theta}^*)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*))$$

Then $\boldsymbol{\theta}^*$ is ε-linearly stable if $\underset{t \to \infty}{lim}\ \mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|] \leq \varepsilon$

Conditions for Linear Stability (Twice Differentiable Minima)

Theorem 1— Linear Stability for SGD in [Wu et al., 2018]

Assume $\nabla \hat{L}_t(\boldsymbol{\theta}^*) = 0$. Then, $\boldsymbol{\theta}^*$ is a linearly stable minimizer if:

$$\lambda_{max}((\mathbb{I} - \eta \nabla^2 \hat{L}_t(\boldsymbol{\theta}^*))^2 + \eta^2 \Sigma) \leq 1$$

Where $\Sigma = \frac{1}{n} \sum_{t=1}^{n} \left[(\nabla^2 \hat{L}_t(\boldsymbol{\theta}^*))^2 - \left(\frac{1}{n} \sum_{t'=1}^{n} \nabla^2 \hat{L}_{t'}(\boldsymbol{\theta}^*) \right)^2 \right]$

- Improvement: Can we relax Assumption on stationery point? $(\nabla \hat{L}_t(\boldsymbol{\theta}^*) = 0 \; \forall \; t \geq 1)$

Theorem 1.1 — Linear Stability for SGD in [Mulayoff et al., 2021]

Consider SGD/GD with step size η, where batches are drawn uniformly from the training set, independently across iterations. If $\boldsymbol{\theta}^*$ is an ε-linearly stable minimum of L, then:

$$\lambda_{max}(\nabla^2 L(\boldsymbol{\theta}^*)) \leq \frac{2}{\eta}$$

Conditions for Linear Stability (Twice Differentiable Minima)

- Assumption 1: Let $\mathbb{E}[\hat{L}_t(\boldsymbol{\theta})] = L(\boldsymbol{\theta})$ and $\mathbb{E}[\nabla \hat{L}_t(\boldsymbol{\theta}^*)] = 0$
- Assumption 2: $\boldsymbol{\theta}^*$ is an ε-linearly stable solution.
- Assumption 3: Batches are drawn uniformly at random, and are independent from each other as $\hat{L}_t(\boldsymbol{\theta})$
- Assumption 4: $\mathbb{E}[\nabla \hat{L}_t(\boldsymbol{\theta}^*)] = 0$

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*] &= \mathbb{E}[\boldsymbol{\theta}_t - \boldsymbol{\theta}^* - \eta\big(\nabla \hat{L}_t(\boldsymbol{\theta}^*) + \nabla^2 \hat{L}_t(\boldsymbol{\theta}^*)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)\big)] && (9)\\
&= \mathbb{E}[\big(\mathbb{I} - \eta \nabla^2 \hat{L}_t(\boldsymbol{\theta}^*)\big)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)] - \eta \mathbb{E}[\nabla \hat{L}_t(\boldsymbol{\theta}^*)] && (10)\\
&= \underbrace{\mathbb{E}[\mathbb{I} - \eta \nabla^2 \hat{L}_t(\boldsymbol{\theta}^*)]\mathbb{E}[\boldsymbol{\theta}_t - \boldsymbol{\theta}^*]}_{\text{Assumption 3}} - \eta \nabla \underbrace{\mathbb{E}[\hat{L}_t(\boldsymbol{\theta}^*)]}_{=0} && (11)\\
&= (\mathbb{I} - \eta \nabla^2 \mathbb{E}[\hat{L}(\boldsymbol{\theta}^*)])\mathbb{E}[\boldsymbol{\theta}_t - \boldsymbol{\theta}^*] && (12)\\
\Rightarrow \|\mathbb{E}[\boldsymbol{\theta}_t - \boldsymbol{\theta}^*]\| &= \|(\mathbb{I} - \eta \nabla^2 L(\boldsymbol{\theta}^*))^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)\| && (13)\\
&\underbrace{\leq \mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|]}_{E[g(X)] \geq g(E[X])} && (14)
\end{aligned}
$$

$$\|\mathbb{E}[\boldsymbol{\theta}_t - \boldsymbol{\theta}^*]\| = \|(\mathbb{I} - \eta\nabla^2 L(\boldsymbol{\theta}^*))^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)\| \tag{15}$$

$$\underbrace{\leq \mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|]}_{E[g(X)]\geq g(E[X])} \tag{16}$$

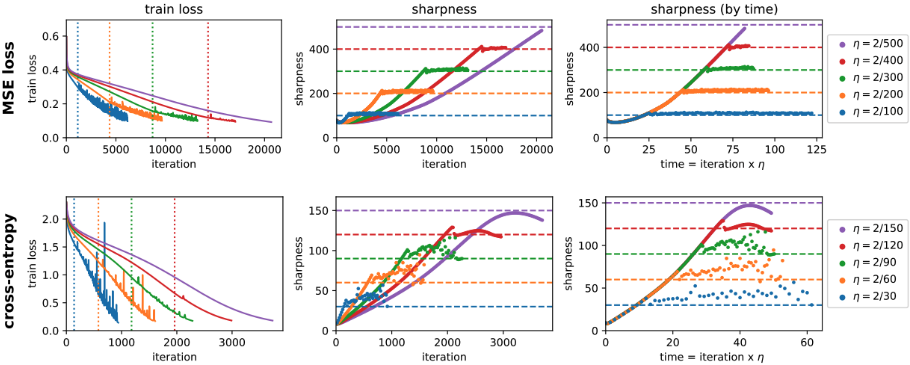
$$\Rightarrow \underset{t\to\infty}{lim}\ \sup\|(\mathbb{I} - \eta\nabla^2 L(\boldsymbol{\theta}^*))^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)\| \leq \varepsilon \quad \text{(Assumption 2)} \tag{17}$$

- Let $\frac{\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*}{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|} = \lambda_{max}(\nabla^2 L(\boldsymbol{\theta}^*))$, and $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\| = \varepsilon$
- Then,

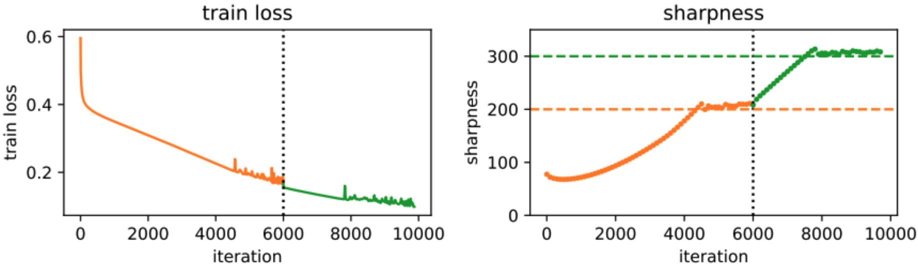
$$\underset{t\to\infty}{lim}\sup|1 - \eta\lambda_{max}(\nabla^2 L(\boldsymbol{\theta}^*))|^t \leq 1 \tag{18}$$

$$\Rightarrow |1 - \eta\lambda_{max}(\nabla^2 L(\boldsymbol{\theta}^*))| \leq 1 \Rightarrow \lambda_{max}(\nabla^2 L(\boldsymbol{\theta}^*)) \leq \frac{2}{\eta} \tag{19}$$

Conditions for Linear Stability: Empirical Study @ICLR2021 [Cohen et al., 2021]



Conditions for Linear Stability: Empirical Study @ICLR2021 [Cohen et al., 2021]



$$\eta = \frac{2}{200} \xrightarrow{\text{6000 Training Iterations}} \eta = \frac{2}{300}$$

Linear Stability (Non-differentiable Minima) (@ NeurIPS '21)

- A more realistic scenario: Non-differentiable minima in deep learning, caused by ReLU activations or max-pooling layers.
- Model SGD's dynamics as a switching dynamical system (SSDS): Fix the activation patterns/ sample
- Let $\{S_m\}$ be a partition of $\mathbb{R}^d$ that represents regions of different modes, $\psi_m : \mathbb{R}^d \to \mathbb{R}$ be a loss function on the m^{th} node. Therefore,

$$L(\boldsymbol{\theta}) = \psi_m(\boldsymbol{\theta}), \qquad \hat{L}_t(\boldsymbol{\theta}) = \hat{\psi}_m^t(\boldsymbol{\theta}) \text{ if } \boldsymbol{\theta} \in S_m$$

$$\forall \boldsymbol{\theta} \in Int(S_m) \quad \hat{g}_{\boldsymbol{\theta}}^t = \nabla \hat{\psi}_m^t(\boldsymbol{\theta}^*) \quad \hat{H}_{\boldsymbol{\theta}}^t = \nabla^2 \hat{\psi}_m^t(\boldsymbol{\theta}^*)$$

- Furthermore, let $\mathcal{I} = \{m : \boldsymbol{\theta}^* \in \bar{S}_m\}$, $\mathcal{A} = \cup_{m \in \mathcal{I}} S_m$

Definition 2— Linear Stability for SGD in for a SSDS:

Assume $\boldsymbol{\theta}^*$ is the minimum of L. Consider the following SSDS:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta(\hat{g}_{\boldsymbol{\theta}_t}^t + \hat{H}_{\boldsymbol{\theta}_t}^t(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*))$$

$\boldsymbol{\theta}^*$ is linearly stable if $\underset{t\to\infty}{lim} \sup \mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|] \le \varepsilon$ for any $\boldsymbol{\theta}_0 \in \mathcal{B}_\varepsilon(\boldsymbol{\theta}^*)$

$\boldsymbol{\theta}^*$ is **linearly-strongly stable** if $\sup_t \mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|] \le \varepsilon$ for any $\boldsymbol{\theta}_0 \in \mathcal{B}_\varepsilon(\boldsymbol{\theta}^*)$

Lemma 3— Linear Stability Condition for SGD in for a SSDS:

(With previous assumptions), Suppose there exists $q \in \mathbb{S}^{d-1}$ and λ_m such that $\|H_m q - \lambda_m q\| \leq \delta$ where $H_m = \nabla^2 \psi_m(\boldsymbol{\theta}^*)$. Then, denote:

$$\lambda^{lower} = \min_{m \in \mathcal{I}} \{\lambda_m\}$$

If

$$\lambda^{lower} > \frac{2}{\eta} + \delta + \frac{\gamma}{\varepsilon}$$

Where $\gamma = \max_{m \in \mathcal{I}} \mathbb{E}[|q^T \hat{g}_m^t|]$, Then $\boldsymbol{\theta}^*$ is not strongly-stable.

What are the properties of minima (in function space) to which SGD converges?

- Consider the set of functions $\mathcal{F}$ that can be implemented by a k-neuron single-layer NN with ReLU activation:

$$\mathcal{F} = \left\{ f : \mathbb{R} \rightarrow \mathbb{R} \middle| f(x) = \sum_{i=1}^{k} w_i^2 \cdot \sigma(w_i^1 x + b_i^1) + b^2 \right\}$$

  With the convex loss function $L(f) = \frac{1}{2n} \sum_{j=1}^{n} (f(x_j - y_j)^2$

- Consider a solution parameter vector:

$$\boldsymbol{\theta} = \left[w_1^{(1)}, \ldots, w_k^{(1)}, b_1^{(1)}, \ldots, b_k^{(1)}, w_1^{(2)}, \ldots, w_k^{(2)}, b^{(2)} \right]$$

- Goal: What are the properties of f in function space, given that we consider f to be accessible by SGD, if there exists some implementation of f that is linearly-stable for SGD.

What are the properties of minima (in function space) to which SGD converges?

- We first compute $\nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ at twice-differentiable global minimum $(f(x_j) = y_j)\ \forall j \in [n]$ (Reasonable assumption in overparam regime)

$$\nabla_{\boldsymbol{\theta}} L = \frac{1}{n} \sum_{j=1}^{n} (f(x_j) - y_j) \nabla_{\boldsymbol{\theta}} f(x_j) \tag{20}$$

Let $\mathcal{I} \in \{0,1\}^k$ be activation of all neurons for input x. Therefore:

$$\begin{cases} [\mathcal{I}(x;\boldsymbol{\theta})]_i = 1 & w_i^{(1)} x + b_i^{(1)} > 0 \\ 0 & otherwise \end{cases}$$

- Then, we can calculate:

$$\nabla_{\boldsymbol{\theta}} f(x) = \begin{bmatrix} \nabla_{w^{(1)}} f(x) \\ \nabla_{b^{(1)}} f(x) \\ \nabla_{w^{(2)}} f(x) \\ \frac{df(x)}{db^2} \end{bmatrix} = \begin{bmatrix} x w^{(2)} \cdot \mathcal{I}(x;\boldsymbol{\theta}) \\ w^{(2)} \cdot \mathcal{I}(x;\boldsymbol{\theta}) \\ \mathcal{I}(x;\boldsymbol{\theta}) \cdot (x w^{(1)} + b^{(1)}) \\ 1 \end{bmatrix}$$

- Let $\Phi = \begin{bmatrix} \nabla_{\boldsymbol{\theta}} f(x_1) & \nabla_{\boldsymbol{\theta}} f(x_2) & \dots & \nabla_{\boldsymbol{\theta}} f(x_n) \end{bmatrix}$

What are the properties of minima (in function space) to which SGD converges?

- Now, calculate the Hessian: $\nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \frac{1}{n}\sum_{j=1}^{n}(\nabla_{\boldsymbol{\theta}} f(x_j))(\nabla_{\boldsymbol{\theta}} f(x_j))^T = \frac{1}{n}\Phi\Phi^T$
- Final Goal: Does an $f \in \mathcal{F}$ have its maximum eigenvalue small enough (from lemma 1 and 2) to allow convergence to f?

$$\Omega(f) = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{3k+1} \mid f(x) = \sum_{i=1}^{k} w_i^{(2)} \sigma\left(w_i^{(1)} x + b_i^{(1)} \right) + b^{(2)} \right\}$$

Lemma 4 — Top Eigenvalue Lower Bound:

Let $f \in \mathcal{F}$ be a twice-differentiable minimizer of the $L_{\boldsymbol{\theta}}(f)$. Then:

$$\min_{\theta \in \Omega(f)} \lambda_{\max}\left(\nabla^2_{\theta}\mathcal{L}\right) \geq 1 + 2\int_{-\infty}^{\infty} \left|f''(x)\right| g(x)\mathrm{d}x$$

Where:

$$g(x) = \begin{cases} \min\left\{g^-(x), g^+(x)\right\}, & x \in [x_{\min}, x_{\max}] \\ 0, & \text{otherwise} \end{cases}$$

Where

$$g^-(x) = \mathbb{P}^2(X < x)\mathbb{E}[x - X \mid X < x]\sqrt{1 + (\mathbb{E}[X \mid X < x])^2}$$
$$g^+(x) = \mathbb{P}^2(X > x)\mathbb{E}[X - x \mid X > x]\sqrt{1 + (\mathbb{E}[X \mid X > x])^2}$$

- Proof? (Appendix Part IV)

- First, we find the maximal eigenvalue of the Hessian in terms of Φ $\lambda_{\max}\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}\right) = \max_{\boldsymbol{v}\in\mathbb{S}^{3k}} \boldsymbol{v}^T\left(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}\right)\boldsymbol{v} = \max_{\boldsymbol{v}\in\mathbb{S}^{3k}} \frac{1}{n}\left\|\Phi^T\boldsymbol{v}\right\|^2 = \max_{\boldsymbol{u}\in\mathbb{S}^{n-1}} \frac{1}{n}\|\Phi\boldsymbol{u}\|^2$
- Take $[\mathcal{I}(x_j,\boldsymbol{\theta})]_i = I_{j,i}$

$$\max_{\boldsymbol{u}\in\mathbb{S}^{n-1}} \frac{1}{n}\|\boldsymbol{\Phi u}\|^2 \geq \frac{1}{n^2}\|\boldsymbol{\Phi 1}\|^2 \qquad \text{(Setting } u=\tfrac{1}{\sqrt{n}}\text{)}$$
$$= 1+\frac{1}{n^2}\sum_{i=1}^k\left[\left(\sum_{j=1}^n x_jI_{j,i}w_i^{(2)}\right)^2+\left(\sum_{j=1}^n I_{j,i}w_i^{(2)}\right)^2+\left(\sum_{j=1}^n\sigma\left(w_i^{(1)}x_j+b_i^{(1)}\right)\right)^2\right]$$
$$= 1+\frac{1}{n^2}\sum_{i=1}^k\left[\left(w_i^{(2)}\right)^2\left(\left(\sum_{j=1}^n x_jI_{j,i}\right)^2+\left(\sum_{j=1}^n I_{j,i}\right)^2\right)+\left(\sum_{j=1}^n\sigma\left(w_i^{(1)}x_j+b_i^{(1)}\right)\right)^2\right)$$
$$\geq 1+\frac{2}{n^2}\sum_{i=1}^k\left|w_i^{(2)}\right|\sqrt{\left(\sum_{j=1}^n x_jI_{j,i}\right)^2+\left(\sum_{j=1}^n I_{j,i}\right)^2}\left|\sum_{j=1}^n\sigma\left(w_i^{(1)}x_j+b_i^{(1)}\right)\right|,$$

- Let $C_i=\{x_j : I_{j,i}=1\}$, $n_i=|C_i|=\sum_{j=1}^n I_{j,i}$

- First, we find the maximal eigenvalue of the Hessian in terms of Φ
- Let $C_i = \{x_j : I_{j,i} = 1\}$, $n_i = |C_i| = \sum_{j=1}^{n} I_{j,i}$

$$
\begin{aligned}
\lambda_{\max}\left(\nabla_\theta^2 \mathcal{L}\right) &\geq 1 + \frac{2}{n^2} \sum_{i=1}^{k} \left|w_i^{(2)}\right| \sqrt{\left(\sum_{x \in C_i} x\right)^2 + n_i^2} \left|\sum_{x \in C_i} \left(w_i^{(1)} x + b_i^{(1)}\right)\right| \\
&= 1 + 2 \sum_{i=1}^{k} \left|w_i^{(2)}\right| \left(\frac{n_i}{n}\right)^2 \sqrt{\left(\frac{1}{n_i} \sum_{x \in C_i} x\right)^2 + 1} \left|\frac{1}{n_i} \sum_{x \in C_i} \left(w_i^{(1)} x + b_i^{(1)}\right)\right| \\
&= 1 + 2 \sum_{i=1}^{k} \left|w_i^{(2)}\right| \left(\mathbb{P}\left(X \in C_i\right)\right)^2 \sqrt{\left(\mathbb{E}\left[X \mid X \in C_i\right]\right)^2 + 1} \left|\mathbb{E}\left[w_i^{(1)} X + b_i^{(1)} \mid X \in C_i\right]\right|
\end{aligned}
$$

- Let $\tau_i = \begin{cases} -\frac{b_i^{(1)}}{w_i^{(1)}}, & w_i^{(1)} \neq 0 \\ 0, & w_i^{(1)} = 0 \end{cases}$

Proof of Lemma 4

- First, we find the maximal eigenvalue of the Hessian in terms of Φ
- Let $C_i = \{x_j : I_{j,i} = 1\}$, $n_i = |C_i| = \sum_{j=1}^{n} I_{j,i}$
- Let $\tau_i = \begin{cases} -\frac{b_i^{(1)}}{w_i^{(1)}}, & w_i^{(1)} \neq 0 \\ 0, & w_i^{(1)} = 0 \end{cases}$
- Then we have:

$$1 + 2\sum_{i=1}^{k} \left|w_i^{(2)}\right| \left(\mathbb{P}\left(X \in C_i\right)\right)^2 \sqrt{\left(\mathbb{E}\left[X \mid X \in C_i\right]\right)^2 + 1} \left|\mathbb{E}\left[w_i^{(1)} X + b_i^{(1)} \mid X \in C_i\right]\right| \geq$$
$$1 + 2\sum_{i=1}^{k} \left|w_i^{(1)} w_i^{(2)}\right| \left(\mathbb{P}\left(X \in C_i\right)\right)^2 \sqrt{\left(\mathbb{E}\left[X \mid X \in C_i\right]\right)^2 + 1} \left|\mathbb{E}\left[X - \tau_i \mid X \in C_i\right]\right|$$

- Also, $\left(\mathbb{P}\left(X \in C_i\right)\right)^2 \sqrt{\left(\mathbb{E}\left[X \mid X \in C_i\right]\right)^2 + 1} \left|\mathbb{E}\left[X - \tau_i \mid X \in C_i\right]\right| \geq min\{g^+(\tau_i), g^-(\tau_i)\}$
- Thus,

$$\lambda_{\max}\left(\nabla_\theta^2 L\right) \geq 1 + 2\sum_{i=1}^{k} \left|w_i^{(1)} w_i^{(2)}\right| \min\left\{g^+\left(\tau_i\right), g^-\left(\tau_i\right)\right\}$$
$$\geq 1 + 2\int_{x_{\min}}^{x_{\max}} \left|f''(x)\right| \min\left\{g^+(x), g^-(x)\right\} \mathrm{d}x, \quad \left\{f''(x) = \sum_{k} w_i^{(1)} w_i^{(2)} \delta(x - \tau_i)\right\}$$
$$\Rightarrow \min_{\theta \in \Omega(f)} \lambda_{\max}\left(\nabla_\theta^2 \mathcal{L}\right) \geq 1 + 2\int_{-\infty}^{\infty} \left|f''(x)\right| g(x) \mathrm{d}x$$

Primary Contribution

Theorem 1.1 — Linear Stability for SGD in [Mulayoff et al., 2021]

Consider SGD/GD with step size η, where batches are drawn uniformly from the training set, independently across iterations. If $\boldsymbol{\theta}^*$ is an ε-linearly stable minimum of L, then:

$$\lambda_{max}(\nabla^2 L(\boldsymbol{\theta}^*)) \leq \frac{2}{\eta}$$

Lemma 4 — Top Eigenvalue Lower Bound:

Let $f \in \mathcal{F}$ be a twice-differentiable minimizer of the $L_{\boldsymbol{\theta}}(f)$. Then:

$$\min_{\theta \in \Omega(f)} \lambda_{\max}\left(\nabla_\theta^2 \mathcal{L}\right) \geq 1 + 2\int_{-\infty}^{\infty} \left|f''(x)\right| g(x) \mathrm{d}x$$

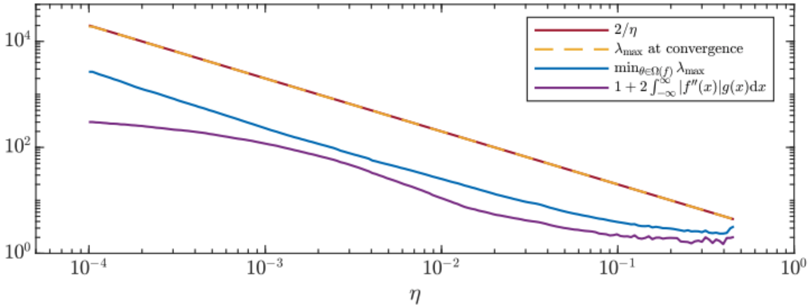
- Clearly, From Lemma 1 and Lemma 4:

$$1 + 2\int_{\mathbb{R}} |f^{''}(x)| g(x) dx \leq \min_{\theta \in \Omega(f)} \lambda_{\max}(\nabla_{\boldsymbol{\theta}}^2 L(\boldsymbol{\theta})) \leq \lambda_{max}(\nabla^2 L(\boldsymbol{\theta}^*)) \leq \frac{2}{\eta}$$

- Clearly, From Lemma 1 and Lemma 4:

$$1 + 2\int_{\mathbb{R}} |f^{''}(x)|g(x)dx \leq \min_{\theta \in \Omega(f)} \lambda_{\max}(\nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta})) \leq \lambda_{max}(\nabla^2 L(\boldsymbol{\theta}^*)) \leq \frac{2}{\eta}$$

- Therefore, $\int_{\mathbb{R}} |f^{''}(x)|g(x)dx \leq \frac{1}{\eta} - \frac{1}{2}$
- Note that similar bound can be constructed for the non-differentiable minima, case.
- Implication: stability in SGD corresponds to the functions with bounded L_1 norm, weighted by a $g(x)$. Furthermore, as we increase learning rate η, smoothness (and flatness) increases.
- Also, bound is initialization independent (no $\boldsymbol{\theta}_0$)



(c) Sharpness versus learning rate

Cohen, J. M., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. (2021).
Gradient descent on neural networks typically occurs at the edge of stability.
arXiv preprint arXiv:2103.00065.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. (2017).
Sharp minima can generalize for deep nets.
In *International Conference on Machine Learning*, pages 1019–1028. PMLR.

Mulayoff, R., Michaeli, T., and Soudry, D. (2021).
The implicit bias of minima stability: A view from function space.
In *Thirty-Fifth Conference on Neural Information Processing Systems.*

Wu, L., Ma, C., et al. (2018).
How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective.
Advances in Neural Information Processing Systems, 31:8279–8288.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016).
Understanding deep learning requires rethinking generalization. arxiv preprint (in iclr 2017).
arXiv preprint arXiv:1611.03530.