Foundations of Deep Learning
School of Computer Science, Tel Aviv University

Instructor: Nadav Cohen
Teaching Assistant: Noam Razin
Scribe: Adi Album & Tomer Epshtein
Date: April 19, 2021

LECTURE
6

Optimization 1

In the course introduction we stated the Three Pillars of Statistical Learning: expressiveness, optimization and generalization. We completed the study of the first chapter - expressiveness. We shall now study optimization, *i.e.* the procedure of updating model's weights in the attempt of minimizing loss on training data.

6.1 Optimization in deep learning is non-convex

In classical ML optimization was quite trivial. The target loss function was convex and therefore optimization schemes are guaranteed to reach global minimum. On the other hand, in DL, any reasonable training program is non-convex. We will prove non-convexity for a general case of settings. Consider a feed-forward fully connected NN with depth $N \geq 2$ and activation $\sigma(\cdot)$:

$$H = \{x \mapsto y = W_N(\sigma(W_{N-1}\sigma(W_{N-2}\sigma(\cdots W_2\sigma(W_1 x))\cdots)|W_n \in \mathbb{R}^{d_n, d_{n-1}}, n \in \mathbb{N}\} \quad (6.1)$$

Proposition 6.1. *Let $L(W_1, W_2, ..., W_N)$ be a loss function that depends on $W_1, W_2, ..., W_N$ only through the input-output mapping of the network. Assume that the global minimum of $L(\cdot)$ is attained for $W_1^*, W_2^*, ..., W_N^*$. Assume also that this global minimum is smaller than the loss attainable with network having hidden widths $d_1 = d_2 = ... = d_{N-1} = 1$. Then $L(\cdot)$ is non-convex.*

Proof. Since $L(\cdot)$ depends on $W_1, W_2, ..., W_N$ only through the input-output mapping, we may permute the rows of W_1^*, and correspondingly permute the columns of W_2^*, such that the value of $L(\cdot)$ is unchanged (still optimal). That is, for any permutation matrix $P \in \mathbb{R}^{d_1, d_1}$: $L(PW_1^*, W_2^* P^T, W_3^*, ..., W_N^*) = L(W_1^*, W_2^*, ..., W_N^*) = L^*$ (optimal loss value).

Assume by contradiction that $L(\cdot)$ is convex. Then:

$$L(\frac{1}{d_1!}\sum_{P\ perm\ mat} PW_1^*, \frac{1}{d_1!}\sum_{P\ perm\ mat} W_2^* P^T, W_3^*, ..., W_N^*) \leq$$
$$\frac{1}{d_1!}\sum_{P\ perm\ mat} L(PW_1^*, W_2^* P^T, W_3^*, ..., W_N^*) = L^*$$

Where $Q := \frac{1}{d_1!}\sum_{P\ perm\ mat} P = \frac{1}{d_1}\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$. We Thus have an optimal weight setting where all the rows of W_1 are the same.
Continuing in this fashion, we can obtain an optimal weight setting $W_1, W_2, ..., W_{N-1}$ whom each have their rows identical (*i.e.* we may write $W_n = \vec{1}\overrightarrow{v_n}^T$ for $n \in [N]$ where $\vec{1}$ is the all-ones vector and $\overrightarrow{v_n} \in \mathbb{R}^{d_{n-1}}$. This contradicts the assumption of L^* being smaller than loss attainable by network with hidden widths $d_1 = d_2 = ... = d_{N-1} = 1$. Thus $L(\cdot)$ is non-convex. □

Note:

- The proposition (6.1) and proof above can easily be extended to the case in which the NN includes biases.
- The result holds for any activation $\sigma(\cdot)$, including linear!

Proposition 6.2. *Let $L(W_1, W_2, ..., W_N)$ be a continuously differentiable loss function that depends on $W_1, W_2, ..., W_N$ only through the input-output mapping of the network. Assume that the global minimum of $L(\cdot)$ is not attained for $W_1 = 0, W_2 = 0, ..., W_N = 0$. Assume also that $\sigma(\cdot)$ is continuously differentiable and $\sigma(0) = 0$. Then, $L(\cdot)$ is non-convex.*

Proof. Homework... □

6.2 Landscape approach

The landscape approach is built on the prmise that if a non-convex objective has no bad local minima and no non-strict saddles, running GD (or SGD) over it will yield global minimum. Such results are typically established using two principles:

- GD reaches stationary (critical) points
- GD escapes strict saddles

6.2.1 Convergence to stationary point

We will show that on any smooth objective, GD with sufficiently small step size converges to (approximate) stationary point in polynomial time.

Definition 6.3. A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if its gradient is β-Lipschitz, *i.e.* :

$$\forall \overrightarrow{w_1}, \overrightarrow{w_2} \in \mathbb{R}^d : \|\nabla f(\overrightarrow{w_1}) - \nabla f(\overrightarrow{w_2})\| \leq \beta \|\overrightarrow{w_1} - \overrightarrow{w_2}\|$$

Lemma 6.4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable and β-smooth. Denote by $\nabla^2 f(\overrightarrow{w})[\cdot,\cdot]$ the bilinear symmetric operator corresponding to the Hessian of $f(\cdot)$ at $\overrightarrow{w}$. Then, for any $\overrightarrow{v} \in \mathbb{R}^d$ it holds that $|\nabla^2 f(\overrightarrow{w})[\overrightarrow{v}, \overrightarrow{v}]| \leq \beta \cdot \|\overrightarrow{v}\|^2$.*

Proof. Define $h : \mathbb{R} \to \mathbb{R}$ by $h(t) = \langle \overrightarrow{v}, \nabla f(\overrightarrow{w} + t\overrightarrow{v}) - \nabla f(\overrightarrow{w}) \rangle$.
From the chain rule we have:

$$h'(0) = \nabla^2 f(\overrightarrow{w})[\overrightarrow{v}, \overrightarrow{v}]$$

On the other hand, by β-smoothness of $f(\cdot)$:

$$\forall t \in \mathbb{R} : |h(t)| \leq \|\overrightarrow{v}\| \cdot \|\nabla f(\overrightarrow{w} + t\overrightarrow{v}) - \nabla f(\overrightarrow{w})\| \leq \|\overrightarrow{v}\| \cdot \beta \cdot \|\overrightarrow{w} + t\overrightarrow{v} - \overrightarrow{w}\| = \beta \cdot t \cdot \|v\|^2$$
$$\implies |h'(0)| \leq \beta \left\|\overrightarrow{v}^2\right\|$$

and the desired result follows. □

Lemma 6.5. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable and β-smooth. Then:*

$$\forall \overrightarrow{w_1}, \overrightarrow{w_2} \in \mathbb{R}^d : |f(\overrightarrow{w_2}) - \underbrace{(f(\overrightarrow{w_1}) + \langle \nabla f(\overrightarrow{w_1}), \overrightarrow{w_2} - \overrightarrow{w_1} \rangle)}_{1^{st} \text{ order approx around } \overrightarrow{w_1}}| \leq \frac{\beta}{2} \|\overrightarrow{w_2} - \overrightarrow{w_1}\|^2$$

Proof. Define $g : \mathbb{R} \to \mathbb{R}$ by $g(t) = f(\overrightarrow{w_1} + t(\overrightarrow{w_2} - \overrightarrow{w_1}))$. By Taylor's theorem:

$$g(1) = g(0) + g'(0) \cdot (1-0) + \frac{1}{2} g''(\xi) \cdot (1-0)^2, \text{ for some } \xi \in (0,1)$$

Plugging in the definition of $g(\cdot)$ and employing the chain rule we get:

$$f(\overrightarrow{w_2}) = f(\overrightarrow{w_1}) + \langle \nabla f(\overrightarrow{w_1}), \overrightarrow{w_2} - \overrightarrow{w_1} \rangle + \frac{1}{2} \nabla^2 f(\overrightarrow{w_1} \cdot (1-\xi) + \overrightarrow{w_2} \cdot \xi)[\overrightarrow{w_2} - \overrightarrow{w_1}, \overrightarrow{w_2} - \overrightarrow{w_1}]$$

The previous lemma (6.4) implies that:

$$|\nabla^2 f(\overrightarrow{w_1} \cdot (1-\xi) + \overrightarrow{w_2} \cdot \xi)[\overrightarrow{w_2} - \overrightarrow{w_1}, \overrightarrow{w_2} - \overrightarrow{w_1}]| \leq \beta \cdot \|\overrightarrow{w_2} - \overrightarrow{w_1}\|^2$$

Using this with the equality above yields the sought after result. □

Definition 6.6. For a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ and $\epsilon \geq 0$, we say that $w \in \mathbb{R}^d$ is an ϵ-stationary point if $\|\nabla f(w)\| \leq \epsilon$
Note: The usual term "Stationary Point" corresponds to the case $\epsilon = 0$

Theorem 6.7. *[1]*
Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice continuously differentiable and β-smooth function that attains its global minimum: $f^* = \min_{\vec{w} \in \mathbb{R}^d} f(\vec{w})$. Suppose we run GD over $f(\cdot)$ with step-size $\eta \leq \frac{1}{\beta}$ starting from $\vec{w_0}$. Then, for any $\epsilon > 0$, an ϵ-stationary point will be reached within a number of steps at most:

$$\frac{2(f(\vec{w_0}) - f^*)}{\eta \cdot \epsilon^2}$$

Proof. For every step $t \geq 0$:

$$\begin{aligned} f(\vec{w_{t+1}}) &\leq f(\vec{w_t}) + \langle \nabla f(\vec{w_t}), \vec{w_{t+1}} - \vec{w_t} \rangle + \frac{\beta}{2} \|\vec{w_{t+1}} - \vec{w_t}\|^2 \\ &\leq f(\vec{w_t}) - \eta \|\nabla f(\vec{w_t})\|^2 + \frac{\beta}{2}\eta^2 \|\nabla f(\vec{w_t})\|^2 \leq f(\vec{w_t}) - \frac{\eta}{2} \|\nabla f(\vec{w_t})\|^2 \end{aligned}$$

Assume that for steps $t = 0, 1, ..., T-1 : \|\nabla f(\vec{w_T})\| > \epsilon$. Then:

$$f(\vec{w_t}) < f(\vec{w_0}) - \frac{\eta}{2} \cdot T \cdot \epsilon^2$$

Since by definition $f(\vec{w_T}) \geq f^*$, we have:

$$f^* < f(\vec{w_0}) - \frac{\eta}{2} \cdot T \cdot \epsilon^2 \Longrightarrow T < \frac{2(f(\vec{w_0}) - f^*)}{\eta \cdot \epsilon^2}$$

Meaning the first step t at which $\|\nabla f(\vec{w_t})\| \leq \epsilon$ comes before step $\frac{2(f(\vec{w_0}) - f^*)}{\eta \cdot \epsilon^2}$ □

6.2.2 Escaping strict saddle point

A strict saddle is a stationary point that has at least one strictly negative eigenvalue to its Hessian (note that this definition can apply to a local maximum). We will consider a 2^{nd} order Taylor approximation around a strict saddle, and show that GD escapes it in reasonable time "with high probability".
Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice continuously differentiable function, with a strict saddle at $\vec{w_s} \in \mathbb{R}^d$. The 2^{nd} order Taylor approximation of $f(\cdot)$ around $\vec{w_s}$ is:

$$\begin{aligned} \tilde{f}(\vec{w}) :&= f(\vec{w_s}) + \langle \nabla f(\vec{w_s}), \vec{w} - \vec{w_s} \rangle + \frac{1}{2}\nabla^2 f(\vec{w_s})[\vec{w} - \vec{w_s}, \vec{w} - \vec{w_s}] \\ &= f(\vec{w_s}) + \frac{1}{2}\nabla^2 f(\vec{w_s})[\vec{w} - \vec{w_s}, \vec{w} - \vec{w_s}] \\ &= f(\vec{w_s}) + \frac{1}{2}(\vec{w} - \vec{w_s})^T H(\vec{w} - \vec{w_s}) \end{aligned}$$

(Where $H \in \mathbb{R}^{d,d}$ is a matrix representing bilinear operator $\nabla^2 f(\vec{w_s})[\cdot, \cdot]$)

Its gradient: $\nabla\tilde{f}(\overrightarrow{w}) = H(\overrightarrow{w} - \overrightarrow{w_s})$, yielding the following dynamics for GD:

$$\overrightarrow{w_{t+1}} = \overrightarrow{w_t} - \eta \cdot H(\overrightarrow{w_t} - \overrightarrow{w_s})$$

For $\epsilon > 0$, we say that GD ϵ-escaped $\overrightarrow{w_s}$ if it reached a non-ϵ-stationary point whose objective value is lower than that of $\overrightarrow{w_s}$, *i.e.* if:

1. $\tilde{f}(\overrightarrow{w_t}) < \tilde{f}(\overrightarrow{w_s})$ $(= f(\overrightarrow{w_s}))$
2. $\left\|\nabla\tilde{f}(\overrightarrow{w_t})\right\| > \epsilon$

Let $H = U\Lambda U^T$ be an orthogonal eigen decomposition for the (symmetric) matrix H, *i.e.* $U \in \mathbb{R}^{d,d}$ is an orthogonal matrix ($UU^T = Id$) and $\Lambda \in \mathbb{R}^{d,d}$ is diagonal. Applying the change of variables $\overrightarrow{w} = U\overrightarrow{\theta} + \overrightarrow{w_s} \iff \overrightarrow{\theta} = U^T(\overrightarrow{w} - \overrightarrow{w_s})$, we have:

- $\tilde{f}(\overrightarrow{w})$ - $f(\overrightarrow{w_s}) = \frac{1}{2}\theta\Lambda\theta^T = \frac{1}{2}\sum_{i=1}^{d}\lambda_i\theta_i^2$ $(\Lambda = diag(\lambda_1, ..., \lambda_d))$
- $\left\|\nabla\tilde{f}(w)\right\| = \|U\Lambda\theta\| = \|\Lambda\theta\| = (\sum_{i=1}^{d}{\lambda_i}^2{\theta_i}^2)^{0.5}$
- $\theta^{(t+1)} = \theta^{(t)}$ - $\eta\Lambda\theta^{(t)} \iff {\theta_i}^{(t+1)} = (1 - \eta\lambda_i){\theta_i}^{(t)}$, $i \in [d]$

Establishing ϵ-escaping thus means showing that under the dynamics of θ_t it holds that $\sum_{i=1}^{d}\lambda_i({\theta_i}^{(t)})^2 < 0$ and $\sum_{i=1}^{d}{\lambda_i}^2({\theta_i}^{(t)})^2 > \epsilon^2$. Plugging in the dynamics, we may write these conditions as follows:

1. $\sum_{i=1}^{d}\lambda_i(1 - \eta\lambda_i)^{2t}\left({\theta_i}^{(0)}\right)^2 < 0$
2. $\sum_{i=1}^{d}{\lambda_i}^2(1 - \eta\lambda_i)^{2t}\left({\theta_i}^{(0)}\right)^2 > \epsilon^2$

If the original objective $f(\cdot)$ is $\beta - smooth$, then by the lemma we've proven all eigenvalues of its Hessian are bounded in absolute value by β. This implies that $\|\lambda_i\| \leq \beta, \forall i \in [d]$. Since $\overrightarrow{w_s}$ is a strict saddle, there exists some $i \in [d]$ for which $\lambda_i < 0$. Assume W.L.O.G that $\lambda_1 = -\alpha < 0$. In the customary setting of step size $\eta \leq \frac{1}{\beta}$:

$\sum_{i=1}^{d}\lambda_i(1 - \eta\lambda_i)^{2t}\left({\theta_i}^{(0)}\right)^2 \leq \lambda_1(1 - \eta\lambda_1)^{2t}\left({\theta_1}^{(0)}\right)^2 + \sum_{i\in[d]:\lambda_i>0}\lambda_i(1 - \eta\lambda_i)^{2t}\left({\theta_i}^{(0)}\right)^2$
$\leq -\alpha(1 + \eta\alpha)^{2t}\left({\theta_1}^{(0)}\right)^2 + \sum_{i\in[d]:\lambda_i>0}\lambda_i\left({\theta_i}^{(0)}\right)^2 < 0$
Where we used $(1 - \eta\lambda_i)^{2t} \in [0, 1]$ and require the last inequality < 0.
$\iff$
$(1 + \eta\alpha)^{2t} > \frac{\sum_{i\in[d]:\lambda_i>0}\lambda_i\left({\theta_i}^{(0)}\right)^2}{({\theta_1}^{(0)})^2\alpha}$
(Assuming ${\theta_1}^{(0)} \neq 0$)
$\iff$
$t > \frac{\frac{1}{2}\log\frac{\sum_{i\in[d]:\lambda_i>0}\lambda_i\left({\theta_i}^{(0)}\right)^2}{({\theta_1}^{(0)})^2\alpha}}{\log(1+\eta\alpha)}$

Additionally:

$$\sum_{i=1}^{d} {\lambda_i}^2(1-\eta\lambda_i)^{2t}({\theta_i}^{(0)})^2 \geq {\lambda_1}^2(1-\eta\lambda_1)^{2t}({\theta_1}^{(0)})^2$$
$$= \alpha^2(1+\eta\alpha)^{2t}({\theta_1}^{(0)})^2 > \epsilon^2$$

Where we require the last inequality $> \epsilon^2$

$$\iff t > \frac{1}{2}\log\left(\epsilon^2/\alpha^2({\theta_1}^{(0)})^2\right)/\log\left(1+\eta\alpha\right)$$
$$\text{(Assuming } {\theta_1}^{(0)} \neq 0)$$

Conditions 1 and 2 thus hold, *i.e.* the saddle is ϵ-escaped, after at most:

$$\left\lceil \frac{1}{2}\frac{log\left(\frac{max\{\sum_{i\in[d]:\lambda_i>0}\lambda_i(\theta_i^{(0)}),\frac{\epsilon^2}{\alpha}\}}{{\theta_1^{(0)}}^2\alpha}\right)}{log(1+\eta\alpha)} \right\rceil \text{ steps.}$$

The reason we said efficient escaping occurs "w.h.p" is because of the dependence on $|{\theta_1}^{(0)}|$ - initial distance from $\overrightarrow{w_s}$ in the direction of negative curvature. The larger this is, the fastest the escape is guaranteed to be. If we initialize $\overrightarrow{w}$ (*i.e.* set $\overrightarrow{w_0}$) by an appropriately scaled isotropic random distribution centered at $\overrightarrow{w_s}$, we can assure that all the coordinates of $\overrightarrow{\theta_0}$ are w.h.p large "enough" (note that there is no need to know which directions have negative curvature).

6.2.3 Putting it all together

Combining the above principles - convergence to stationary point and escaping non-strict saddle point - into a formal result (guarantee of efficient convergence to stationary point with PSD Hessian) is a highly technical process, primarily since one needs to treat the discrepancy between 2^{nd} order Taylor approx and the original objective. We will not do this here. Instead, we will state an exemplar result, taken from [2].

Definition 6.8. A twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is a ρ-Hessian-Lipschitz if:

$$\forall \overrightarrow{w_1}, \overrightarrow{w_2} \in \mathbb{R}^d : \left\|\nabla^2 f(\overrightarrow{w_1}) - \nabla^2 f(\overrightarrow{w_2})\right\|_{Spectral} \leq \rho \left\|\overrightarrow{w_1} - \overrightarrow{w_2}\right\|$$

For such function, we say that $\overrightarrow{w}$ is an $\epsilon - 2^{nd} - order\ stationary\ point$ if:

$$\|\nabla f(\overrightarrow{w})\| \leq \epsilon$$

and

$$\lambda_{min}(\nabla^2 f(\overrightarrow{w})) \geq -\sqrt{\rho\epsilon}$$

Algorithm: Perturbed Gradient Descent: PGD$(\overrightarrow{w_0}, \beta, \rho, \epsilon, c, \delta, \Delta f)$:

1. $x \leftarrow 3\max\{\log(\frac{d\beta\Delta f}{c\epsilon^2\delta}), 4\}$, $\eta \leftarrow \frac{c}{\beta}$, $r \leftarrow \frac{\sqrt{c}}{x^2} \cdot \frac{\epsilon}{\beta}$
2. $g_{thresh} \leftarrow \frac{\sqrt{c}}{x^2} \cdot \epsilon$, $f_{thresh} \leftarrow \frac{c}{x^3} \cdot \sqrt{\frac{\epsilon^3}{\rho}}$, $t_{thresh} \leftarrow \frac{x}{c^2} \cdot \frac{\beta}{\sqrt{\rho\epsilon}}$
3. $t_{noise} \leftarrow -t_{thresh} - 1$
4. for $t = 0, 1, ..., d_0$:
   - 4.1. If $\|\nabla f(\overrightarrow{w_t})\| \leq g_{thresh}$ and $t - t_{noise} > t_{thresh}$ then:
     - 4.1.1. $\overrightarrow{\tilde{w}_t} \leftarrow \overrightarrow{w_t}$, $t_{noise} \leftarrow t$
     - 4.1.2. $\overrightarrow{w_t} \leftarrow \overrightarrow{\tilde{w}_t} + \overrightarrow{\xi_t}$, where $\overrightarrow{\xi_t}$ is uniformly sampled over $\{\overrightarrow{w'} : \left\|\overrightarrow{w'}\right\| \leq r\}$
   - 4.2. If $t - t_{noise} = t_{thresh}$ and $f(\overrightarrow{w_t}) - f(\overrightarrow{\tilde{w}_{t_{noise}}}) > -f_{thresh}$ then:
     - 4.2.1. return $\overrightarrow{\tilde{w}_{t_{noise}}}$
   - 4.3. $\overrightarrow{w_{t+1}} \leftarrow \overrightarrow{w_t} - \eta \cdot \nabla f(\overrightarrow{w_t})$

Note: Perturbation ξ_t in (4.1.2.) is needed in order to ensure sufficient distance from saddle in direction of negative curvature.

Theorem 6.9. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a β-smooth and ρ-Hessian Lipschitz function with global minimum f^*. Then there exists an absolute constant c_{max} s.t. for any $\delta > 0$, $\epsilon \leq \frac{\beta^2}{\rho}$, $\Delta f \geq f(\overrightarrow{w_0}) - f^*$ and const $c \leq c_{max}$, PGD$(\overrightarrow{w_0}, \beta, \rho, \epsilon, c, \delta, \Delta f)$ will output an ϵ-2^{nd}-order stationary point, w.p. $1 - \delta$, and terminate in the following number of steps:*

$$\mathcal{O}\left(\frac{\beta(f(\overrightarrow{w_0}) - f^*)}{\epsilon^2} \cdot \log^4(\frac{d\beta\Delta f}{\epsilon^2\delta})\right)$$

6.2.4 Example - linear neural networks

As a simple test-bed for the landscape approach, we will analyze linear neural networks (LNN) - feed-forward fully connected NN with linear (no) activation. Note that in accordance with what wa've shown above, they induce non-convex training programs, even though their input-output mappings are linear.
A depth N LNN with input dim $d_0 \in \mathbb{N}$, hidden dims $d_1, d_2, ..., d_{N-1} \in \mathbb{N}$, and output dim $d_N \in \mathbb{N}$, is the following parametric family of functions:

$$\{x \in \mathbb{R}^{d_0} \mapsto y = W_N W_{N-1} \cdots W_1 x \in \mathbb{R}^{d_N} : W_n \in \mathbb{R}^{d_n, d_{n-1}}, n \in [N]\}$$

Given a LNN, for $1 \leq j \leq j' \leq N$ we denote: $W_{j:j'} := W_{j'} W_{j'-1} \cdots W_j$. For convenience, if $j > j'$, $W_{j:j'}$ stands for identity with size to be inferred from context.
Let $l : \mathbb{R}^{d_N, d_0} \to \mathbb{R}$ be a twice continuously differentialbe convex loss (*e.g.* logistic regression over linear models). Denote its gradient at $W \in \mathbb{R}^{d_N, d_0}$ by $\nabla l(W) \in \mathbb{R}^{d_N, d_0}$, and the

bilinear symmetric (PSD) operator associated with its Hessian (at W) by $\nabla^2 l(W)[\cdot,\cdot]$: $\mathbb{R}^{d_N,d_0} \times \mathbb{R}^{d_N,d_0} \mapsto \mathbb{R}$. $l(\cdot)$ induces an overparameterized objective for the LNN:

$$\begin{aligned}&\Phi : \mathbb{R}^{d_1,d_0} \times \mathbb{R}^{d_2,d_1} \times \cdots \times \mathbb{R}^{d_N,d_{N-1}} \to \mathbb{R} \\ &\Phi(W_1, W_2, ..., W_N) = l(W_{1:N}) := l(W_N W_{N-1} \cdots W_1)\end{aligned} \tag{6.2}$$

We will analyze the (non-convex) landscape of $\Phi(\cdot)$

6.2.5 Gradient and Hessian

We start by computing the gradient and Hessian of $\Phi(\cdot)$. For any $\Delta_j \in \mathbb{R}^{d_j,d_{j-1}}, j \in [N]$:

$$\begin{aligned}&\Phi(W_1 + \Delta_1, W_2 + \Delta_2, ..., W_N + \Delta_N) = \\ &= l\left((W_N + \Delta_N)(W_{N-1} + \Delta_{N-1}) \cdots (W_1 + \Delta_1)\right) = \\ &= l\left(W_{1:N} + \sum_{j=1}^{N} (W_{j+1:N} \Delta_j W_{1:j-1}) + \sum_{1 \le j \le j' \le N} (W_{j'+1:N} \Delta_{j'} W_{j+1:j'-1} \Delta_j W_{1:j-1}) + \mathbf{o}(\|(\Delta_1, ..., \Delta_N)\|_{Fro}^2)\right)\end{aligned}$$

Where little $\mathbf{o}$: $\lim_{(\Delta_1,...,\Delta_N)\to 0} \frac{\mathbf{o}(\|(\Delta_1,...,\Delta_N)\|_{Fro}^2)}{\|(\Delta_1,...,\Delta_N)\|_{Fro}^2} = 0$

Plugging in the 2^{nd}-order Taylor expansion of $l(\cdot)$:

$$l(W + \Delta) = l(W) + \langle \nabla l(W), \Delta \rangle + \frac{1}{2} \nabla^2 l(W)[\Delta, \Delta] + \mathbf{o}(\|\Delta\|_{Fro}^2)$$

We obtain:

$$\begin{aligned}\Phi(W_1 + \Delta_1, W_2 + \Delta_2, ..., W_N + \Delta_N) = l(W_{1:N}) + \langle \nabla l(W_{1:N}), \sum_{j=1}^{N} W_{j+1:N} \Delta_j W_{1:j-1} \rangle + \\ \langle \nabla l(W_{1:N}), \sum_{1 \le j \le j' \le N} (W_{j'+1:N} \Delta_{j'} W_{j+1:j'-1} \Delta_j W_{1:j-1}) \rangle + \\ \frac{1}{2} \nabla^2 l(W_{1:N})[\sum_{j=1}^{N} W_{j+1:N} \Delta_j W_{1:j-1}, \sum_{j=1}^{N} W_{j+1:N} \Delta_j W_{1:j-1}] + \\ \mathbf{o}(\|(\Delta_1, ..., \Delta_N)\|_{Fro}^2)\end{aligned}$$

This is in fact a 2^{nd}-order Taylor expansion of $\Phi(\cdot)$ around $(W_1, ..., W_N)$:

- 0^{th}-order term:

$$\Phi(W_1, ..., W_N) = l(W_{1:N})$$

- 1^{st}-order term (gradient):

$$\begin{aligned}&\langle \nabla \Phi(W_1, ..., W_N), (\Delta_1, ..., \Delta_N) \rangle = \langle \nabla l(W_{1:N}), \sum_{j=1}^{N} W_{j+1:N} \Delta_j W_{1:j-1} \rangle \\ \implies &\nabla \Phi(W_1, ..., W_N) = ({W_{2:N}}^T \nabla l(W_{1:N}), ..., {W_{j+1:N}}^T \nabla l(W_{1:N}) {W_{1:j-1}}^T, ..., \nabla l(W_{1:N}) {W_{1:N-1}}^T)\end{aligned}$$

- 2^{nd}-order term (Hessian):

$$\nabla^2\Phi(W_1,...,W_N)[(\Delta_1,...\Delta_N),(\Delta_1,...\Delta_N)] = \nabla^2 l(W_{1:N})[\sum_{j=1}^{N} W_{j+1:N}\Delta_j W_{1:j-1}, \sum_{j=1}^{N} W_{j+1:N}\Delta_j W_{1:j-1}]$$
$$+2\langle \nabla l(W_{1:N}), \sum_{1\le j\le j'\le N} (W_{j'+1:N}\Delta_{j'} W_{j+1:j'-1}\Delta_j W_{1:j-1})\rangle$$

6.2.6 No bad local minima

We will show that $\Phi(\cdot)$ has no bad local minina. The proof we employ is taken from [3] and assumes no bottleneck layers, *i.e.* $min\{d_1,...d_{N-1}\} \ge min\{d_0,d_N\}$. There are many other proofs in the literature which make different assumptions - we choose this for its simplicity.

Theorem 6.10. *let $l : \mathbb{R}^{d_N,d_0} \to \mathbb{R}$ be a differentiable convex function including an overparameterized objective $\Phi(\cdot)$ via (6.2). Assume that $min\{d_1,...d_{N-1}\} \ge min\{d_0,d_N\}$. Then, any local minimizer $(\widehat{W}_1,...,\widehat{W}_N)$ of $\Phi(\cdot)$ is a global minimizer.*

Proof. Assume w.l.o.g that $d_N \ge d_0$. (If this is not the case the we consider $\tilde{l} : \mathbb{R}^{d_0,d_N} \to \mathbb{R}$ defined by $\tilde{l}(W) = l(W^T)$. Then $d_n \ge d_0\ \forall n \in [N]$. Let $(\widehat{W}_1,\widehat{W}_2,...,\widehat{W}_N)$ be a local minimizer of $\Phi(\cdot)$. 1^{st} order optimality condition implies:

$$\frac{\partial}{\partial W_N}\Phi(\widehat{W}_1,\widehat{W}_2,...,\widehat{W}_N) = \nabla l(\widehat{W}_{1:N})\widehat{W}_{1:N-1}^T = 0$$

If $\widehat{W}_{1:N-1} \in \mathbb{R}^{d_{N-1},d_0}$ has a trivial kernel (*i.e.* has linearly dependant columns) then necessarily $\nabla l(\widehat{W}_{1:N}) = 0$, which by convexity implies that $\widehat{W}_{1:N}$ is a global min of $l(\cdot)$, *i.e.* $(\widehat{W}_1,...,\widehat{W}_N)$ is a global min for $\Phi(\cdot)$. We now assume that $\widehat{W}_{1:N-1}$ has a non-trivial kernel. It holds that:

$$Ker(\widehat{W}_{1:1}) \subseteq Ker(\widehat{W}_{1:2}) \subseteq \cdots \subseteq Ker(\widehat{W}_{1:N-1})$$

thus there exists a $k^* \in [N-1]$ s.t.:

$$Ker(\widehat{W}_{1:k}) = \{0\}, \forall k < k^*$$

and

$$Ker(\widehat{W}_{1:k}) \ne \{0\}, \forall k \ge k^*$$

For $k \geq k^*$, consider the SVD of $\widehat{W_{1:k}}$:

$$\widehat{W}_{1:k} = \widehat{U}_k \widehat{\Sigma}_k \widehat{V}_k^T$$

Where $\widehat{U}_k \in \mathbb{R}^{d_k, d_k}$ and $\widehat{V}_k \in \mathbb{R}^{d_0, d_0}$ are orthogonal matrices, $\widehat{\Sigma}_{k,k} \in \mathbb{R}^{d_k, d_0}$ holds $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_{d_0} \geq 0$ on its diagonal and zeros elsewhere.

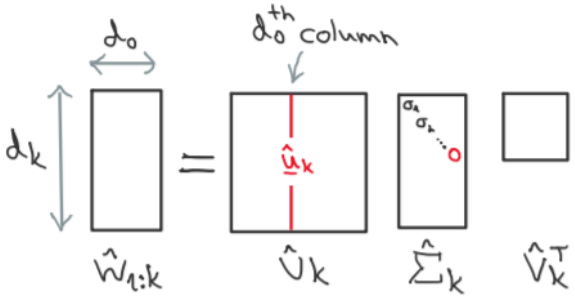

Figure 6.1: SVD decomposition

Since $Ker(\widehat{W}_{1:k}) \neq \{0\}$ it must hold that $\sigma_{d_0} = 0$. Denote by $\overrightarrow{\widehat{u}_k}$ the corresponding (*i.e.* the d_0^{th}) column of $\widehat{U}_k$. Let $\{\overrightarrow{w_k} \in \mathbb{R}^{d_k}\}_{k=k^*+1}^N$ be an arbitrary collection of vectors, and define $(\widetilde{W}_1, \widetilde{W}_2, ..., \widetilde{W}_N)$ by:

$$\widetilde{W}_k := \begin{cases} \widehat{W}_k & \text{if } k \leq k^* \\ \widehat{W}_k + \overrightarrow{w_k}\overrightarrow{\widehat{u}_{k-1}} & k > k^* \end{cases}$$

We claim that $\widetilde{W}_{1:N} = \widehat{W}_{1:N}$. (Reminder: $\widetilde{W}_{j:j'} := \widetilde{W}_{j'}\widetilde{W}_{j'-1} \cdots \widetilde{W}_j$).
To see this, note that $\widetilde{W}_{1:k^*} = \widehat{W}_{1:k^*}$ holds trivially, and by induction over $k \geq k^*$, if $\widetilde{W}_{1:k} = \widehat{W}_{1:k}$ then:

$$\widetilde{W}_{1:k+1} = \widetilde{W}_{k+1}\widetilde{W}_{1:k} = (\widehat{W}_{k+1} + \overrightarrow{w_{k+1}}\overrightarrow{\widehat{u}_k}^T)\widehat{W}_{1:k} = \widehat{W}_{1:k+1} + \overrightarrow{w_{k+1}}\overbrace{\overrightarrow{\widehat{u}_k}^T \widehat{U}_k \widehat{\Sigma}_k}^{=0}\widehat{V}_k^T = \widehat{W}_{1:k+1}$$

So, indeed $\widetilde{W}_{1:N} = \widehat{W}_{1:N}$, meaning that $\Phi(\widetilde{W}_1, \widetilde{W}_2, ..., \widetilde{W}_N) = \Phi(\widehat{W}_1, \widehat{W}_2, ..., \widehat{W}_N)$. By taking small $\{\overrightarrow{w_k}\}_{k=k^*+1}^N$, $(\widetilde{W}_1, \widetilde{W}_2, ..., \widetilde{W}_N)$ can be made arbitrarily close to $(\widehat{W}_1, \widehat{W}_2, ..., \widehat{W}_N)$. In particular, since $(\widehat{W}_1, \widehat{W}_2, ..., \widehat{W}_N)$ is a local min (of $\Phi(\cdot)$), there exists $\epsilon > 0$ s.t. if $\|\overrightarrow{w_k}\| \leq \epsilon$ for all $k^* + 1 \leq k \leq N$, $(\widetilde{W}_1, \widetilde{W}_2, ..., \widetilde{W}_N)$ will be a local min as well. We can thus apply 1^{st} order optimality condition for this case:

$$\widetilde{W}_{j+1:N}^T \nabla l(\widetilde{W}_{1:N}) \widetilde{W}_{1:j-1}^T = 0, \forall j \in [N]$$

In particular we have:

$$0 = \widetilde{W}_{k^*+1:N}^T \nabla l(\widetilde{W}_{1:N}) \widetilde{W}_{1:k^*-1}^T = \widetilde{W}_{k^*+1:N}^T \nabla l(\widehat{W}_{1:N}) \widetilde{W}_{1:k^*-1}^T$$

(last equality holds because $\widetilde{W}_{1:N} = \widehat{W}_{1:N}$ and $\widetilde{W}_k = \widehat{W}_k$, $\forall k \in [k^*]$)

$$\implies \widehat{W}_{1:k^*-1} \nabla l(\widehat{W}_{1:N})^T \widetilde{W}_{k^*+1:N} = 0$$

By definition of k^*: $Ker(\widehat{W}_{1:k^*-1}) = 0$, so:

$$\nabla l(\widehat{W}_{1:N})^T \widetilde{W}_{k^*+1:N} = 0$$

By definition of $\widetilde{W}_{k^*+1}$, we have:

$$\nabla l(\widehat{W}_{1:N})^T \widetilde{W}_{k^*+2:N} (\widehat{W}_{k^*+1} + \overrightarrow{w_{k^*+1}} \overrightarrow{\widehat{u}_{k^*}}^T) = 0$$

Recall that this holds for any choise of $\{\overrightarrow{w_k}\}_{k=k^*+1}^N$ s.t. $\|\overrightarrow{w_k}\| \leq \epsilon$, $\forall k$. Subtracting from the latter equation, obtained when replacing $\overrightarrow{w_{k^*+1}}$ by the zero vector, we get:

$$\nabla l(\widehat{W}_{1:N})^T \widetilde{W}_{k^*+2:N} \overrightarrow{w_{k^*+1}} \overrightarrow{\widehat{u}_{k^*}}^T = 0$$

$$\implies \nabla l(\widehat{W}_{1:N})^T \widetilde{W}_{k^*+2:N} \overrightarrow{w_{k^*+1}} \overbrace{\overrightarrow{\widehat{u}_{k^*}}^T \overrightarrow{\widehat{u}_{k^*}}}^{=1} = \nabla l(\widehat{W}_{1:N})^T \widetilde{W}_{k^*+2:N} \overrightarrow{w_{k^*+1}} = 0$$

Since $\overrightarrow{w_{k^*+1}}$ is an arbitrary (sufficiently small) vector, necessarily $\nabla l(\widehat{W}_{1:N})^T \widetilde{W}_{k^*+2:N} = 0$. Continuing in this fashion leads to $\nabla l(\widehat{W}_{1:N}) = 0$, which by convexity of $l(\cdot)$ implies that $(\widehat{W}_1, \widehat{W}_2, ..., \widehat{W}_N)$ is a global min for $\Phi(\cdot)$, as required. □

6.2.7 No non-strict saddles?

Are all saddle points in $\Phi(\cdot)$ strict? We start by showing that for depth $N = 2$ this is indeed the case. The proof we employ is adopted from [4], and assumes square matrices, *i.e.* $d_0 = d_1 = d_2$. There are many proofs in the literature which make different assumptions - We chose this for its simplicity.

Theorem 6.11. *Let $l : \mathbb{R}^{d,d} \to \mathbb{R}$ be a twice continuously differentiable convex loss. Consider a LNN of depth $N = 2$ and hidden width $d_1 = d$, and let $\Phi(\cdot)$ be the overparameterized objective defined by (6.2). Then, any stationary point $(\widehat{W}_1, \widehat{W}_2)$ of $\Phi(\cdot)$ which is not a global minimizer is a strict saddle.*

Proof. For $(\widehat{W}_1, \widehat{W}_2)$ a stationary point:

$$\nabla \Phi(\widehat{W}_1, \widehat{W}_2) = (\widehat{W}_2^T \nabla l(\widehat{W}_{1:2}^T), \nabla l(\widehat{W}_{1:2}^T \widehat{W}_1^T)) = (0, 0)$$

Assume that $(\widehat{W}_1, \widehat{W}_2)$ in not a global min of $\Phi(\cdot)$. Then $\widehat{W}_{1:2}$ is not a global min of $l(\cdot)$, which by convexity means $\nabla l(\widehat{W}_{1:2}) \neq 0$. This implies:

- There exists $(i,j) \in [d]X[d]$ s.t $\nabla l(\widehat{W}_{1:2})_{i,j} = c \neq 0$
- $\widehat{W}_1$ and $\widehat{W}_2$ are singular (otherwise we obtain a contradiction to $\nabla\Phi(\widehat{W}_1, \widehat{W}_2) = (0,0)$

Let $\overrightarrow{0} \neq \overrightarrow{v} \in \mathbb{R}^d$ be s.t $W_2\overrightarrow{v} = \overrightarrow{0}$, and for $\alpha \in \mathbb{R}$ define:

$$\Delta_1 = \alpha \cdot \overrightarrow{v}\overrightarrow{e_j}^T \in \mathbb{R}^{d,d}, \Delta_2 = \overrightarrow{e_i}\overrightarrow{v}^T \in \mathbb{R}^{d,d}$$

where $\overrightarrow{e_i}$, $\overrightarrow{e_j}$ indicator vectors (vectors with one in location marked by sub index, and zero elsewhere)

From our derivation of $\nabla^2\Phi(\cdot)$:

$$\nabla^2\Phi(\widehat{W}_1, \widehat{W}_2)[(\Delta_1,\Delta_2),(\Delta_1,\Delta_2)] = \nabla^2 l(\widehat{W_{1:2}})[\widehat{W}_2\Delta_1 + \Delta_2\widehat{W}_1, \widehat{W}_2\Delta_1 + \Delta_2\widehat{W}_1] \\ +2\langle\nabla l(\widehat{W}_{1:2}), \Delta_2\Delta_1\rangle$$

It holds that:

- $\widehat{W}_2\Delta_1 = \alpha \cdot \widehat{W}_2 \cdot \overrightarrow{v} \cdot \overrightarrow{e_j}^T = 0$ because $\widehat{W}_2\overrightarrow{v} = \overrightarrow{0}$
- $\Delta_2\Delta_1 = \alpha \cdot \overrightarrow{e_i} \cdot \overrightarrow{v}^T \cdot \overrightarrow{v} \cdot \overrightarrow{e_j}^T = \alpha\|v\|^2 \cdot \overrightarrow{e_i} \cdot \overrightarrow{e_j}^T$

And thus:

$$\nabla^2\Phi(\widehat{W}_1, \widehat{W}_2)[(\Delta_1,\Delta_2),(\Delta_1,\Delta_2)] = \nabla^2\Phi(\widehat{W}_{1:2})[\Delta_2\widehat{W}_1, \Delta_2\widehat{W}_1] + \alpha \cdot 2\|v\|^2 \cdot \langle\nabla^2 l(\widehat{W}_{1:2}), \overrightarrow{e_i} \cdot \overrightarrow{e_j}^T\rangle$$

Note that the left term in the sum does not depend on α, and
$\|v\|^2 \neq 0$ and $\langle\nabla^2 l(\widehat{W}_{1:2}), \overrightarrow{e_i} \cdot \overrightarrow{e_j}^T\rangle = c \neq 0$
Thus implies that we can choose $\alpha \in \mathbb{R}$ s.t $\nabla^2\Phi(\widehat{W}_1, \widehat{W}_2)[\Delta_2\widehat{W}_1, \Delta_2\widehat{W}_1] < 0$. Hence, $\nabla^2\Phi(\widehat{W}_1, \widehat{W}_2)$ has negative eigenvalues, meaning $(\widehat{W}_1, \widehat{W}_2)$ is a strict saddle, as required. □

Moving on to depth $N \geq 3$.

Proposition 6.12. *Let $l : \mathbb{R}^{d_N,d_0} \to \mathbb{R}$ be a twice continuously differentiable convex function. Consider a LNN of depth $N \geq 3$, with hidden widths $d_1, ...d_{N-1}$ s.t $min(d_1, ..., d_{N-1}) \geq min(d_0, d_N)$, and let $\Phi(\cdot)$ be the overparameterized objective defined by (6.2). Assume that $l(\cdot)$ does not attain its global min at 0. Then, $\Phi(\cdot)$ has non-strict saddles.*

Proof. Consider the point $\widehat{W}_1 = 0, ..., \widehat{W}_N = 0$. By assumption $\widehat{W}_{1:N} = \widehat{W}_N \cdots \widehat{W}_1$ is not a global min of $l(\cdot)$, thus $(\widehat{W}_1, ..., \widehat{W}_N)$ is not a global min of $\Phi(\cdot)$. The latter has no bad local minima so $(\widehat{W}_1, ..., \widehat{W}_N)$ is not a local minimum either. On the other hand, by our derivations above, it is clear that $\nabla\Phi(\widehat{W}_1, ..., \widehat{W}_N) = (0, ..., 0)$ and $\nabla^2\Phi(\widehat{W}_1, ..., \widehat{W}_N)[\cdot,\cdot]$ is the zero operator (all eigenvalues = 0). $(\widehat{W}_1, ..., \widehat{W}_N)$ is thus a non-strict saddle. □

We conclude that even with the simplest models (LNNs), the no non-strict saddle property is violated when depth is greater than two. The landscape approach (in its current form) is thus unsuitable for establishing convergence of GD to global min over deep NNs. A different perspective is needed...

References

[1] Yu. Nesterov. Introductory Lectures on Convex Programming. In *Volume I: Basic course*, 1998.

[2] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to Escape Saddle Points Efficiently. In *International Conference on Machine Learning, PMLR*, 2017.

[3] Thomas Laurent and James von Brecht. Deep linear neural networks with arbitrary loss: All local minima are global. In *International Conference on Machine Learning, PMLR*, 2017.

[4] Maher Nouiehed and Meisam Razaviyayn. Learning Deep Models: Critical Points and Local Openness. In *ICLR*, 2018.