# An Optimization and Generalization Analysis for Max-Pooling Networks
## *Supplementary Material*

**Alon Brutzkus**[1] **Amir Globerson**[1]

[1]The Blavatnik School of Computer Science, Tel Aviv University

## A CONVERGENCE RATES FOR THEOREM 5.1

In Ji and Telgarsky [2019], Theorem 4.2, they show the following for logistic regression initialized at zero and a certain learning rate schedule. The margin of the learned classifier is $\frac{\gamma}{2}$ where $\gamma$ is the max-margin after $O\left(\frac{1}{\gamma^2}\right)$ iterations.[1] They show this for normalized points with norm 1. In our case (see the proof of Theorem 5.1), the max margin after normalizing the points to have norm 1, is $\frac{1}{\sqrt{d}}$. Thus, under their assumptions, after $O(d)$ iterations we converge to a solution whose margin is a $\frac{1}{2}$-multiplicative approximation of the max margin. Therefore, we obtain for this solution, up to a constant, the same generalization guarantees as the max margin classifier (which we provide in the theorem).

## B PROOF OF LEMMA 5.2

By definition of the initialization we have $\mathbb{P}\left(i \in \mathcal{A}^+\right) = \frac{1}{2}$. Furthermore, we have that $\mathbb{P}\left(i \in \mathcal{W}_0^+\right) = \frac{\left(1-2^{-d+1}\right)}{d-1}$. This follows, since with probability $2^{-d+1}$, for all $\boldsymbol{o} \in \mathcal{O} \setminus \{2\}$, $\boldsymbol{w}_i^{(0)} \cdot \boldsymbol{o} \le 0$. On the other hand, with probability $\left(1 - 2^{-d+1}\right)$, there exists at least one $\boldsymbol{o} \in \mathcal{O} \setminus \{2\}$ such that $\boldsymbol{w}_i^{(0)} \cdot \boldsymbol{o} > 0$. Assume we condition on the latter event. Then, we get by symmetry that $\boldsymbol{o}_1$ maximizes the dot product with $\boldsymbol{w}_i^{(0)}$, among patterns in $\mathcal{O} \setminus \{2\}$, with probability $\frac{1}{d-1}$.

By independence of $W_0$ and $\boldsymbol{a}^{(0)}$, we have: $\mathbb{P}\left(i \in \mathcal{W}_0^+ \cap \mathcal{A}^+\right) = \frac{\left(1-2^{-d+1}\right)}{2(d-1)}$. Then, by Hoeffding's inequality we get:

$$
\begin{aligned}
\mathbb{P}\left(\left|\frac{|\mathcal{W}_0^+ \cap \mathcal{A}^+|}{k} - \frac{\left(1-2^{-d+1}\right)}{2(d-1)}\right| > \frac{1}{4d}\right) &\le 2e^{-2k\left(\frac{1}{4d}\right)^2} \\
&\le 2e^{-d} \qquad (1)
\end{aligned}
$$

where in the last inequality we used the assumption on $k$. Since $\frac{\left(1-2^{-d+1}\right)}{2(d-1)} \ge \frac{1}{2d}$ and $\frac{\left(1-2^{-d+1}\right)}{2(d-1)} \le \frac{1}{d}$ for $d \ge 3$, we get that with probability at least $1 - 2e^{-d}$, $|\mathcal{W}_0^+ \cap \mathcal{A}^+| \ge \frac{\left(1-2^{-d+1}\right)k}{2(d-1)} - \frac{k}{4d} \ge \frac{k}{4d}$ and $|\mathcal{W}_0^+ \cap \mathcal{A}^+| \le \frac{\left(1-2^{-d+1}\right)k}{2(d-1)} + \frac{k}{4d} \le \frac{k}{d}$. By the symmetry of our problem and definitions of the sets $\mathcal{W}_0^+$, $\mathcal{W}_0^-$, $\mathcal{A}^+$, $\mathcal{A}^-$, we similarly get that with probability at least $1 - 2e^{-d}$, $\frac{k}{4d} \le |\mathcal{W}_0^- \cap \mathcal{A}^-| \le \frac{k}{d}$. Applying the union bound concludes the proof.

## C PROOF OF LEMMA 5.3

We first prove the following two auxiliary lemmas.

**Lemma C.1.** *For all $0 \le t \le T_1$ and all $1 \le i \le k$,* $\left\|\boldsymbol{w}_i^{(t)}\right\| \le \eta_1(t+1)$.

[1]$O$ hides a dependency on $\log m$.

*Proof.* First we notice that for all $1 \le i \le k$, $\left\|\frac{\partial \mathcal{L}_1}{\partial \boldsymbol{w}_i}\left(W, \boldsymbol{a}^{(0)}\right)\right\| \le 1$. This follows since for all $1 \le j \le n$ and all $\boldsymbol{x} \in S_1$, $\|\boldsymbol{x}[j]\| = 1$ (recall that $\|\boldsymbol{o}\| = 1$ for $\boldsymbol{o} \in \mathcal{O}$).

Therefore, for all $0 \le t \le T_1$ and $1 \le i \le k$, $\left\|\boldsymbol{w}_i^{(t)}\right\| \le r + \eta_1 t \le \eta_1(t+1)$. □

**Lemma C.2.** *For all $\boldsymbol{x} \in S_1$ and $0 \le t \le T_1$ $\left|N_{\text{CNN}}\left(\boldsymbol{x}; \left(W^{(t)}, a^{(0)}\right)\right)\right| \le \frac{1}{2}$.*

*Proof.* By Lemma C.1 we have for all $\boldsymbol{x} \in S_1$:

$$\begin{aligned}\left|N_{\text{CNN}}\left(\boldsymbol{x}; \left(W^{(t)}, a^{(0)}\right)\right)\right| &= \left|\sum_{i=1}^{k} a_i^{(0)}\left[\max_j \left\{\sigma\left(\boldsymbol{w}_i^{(t)} \cdot \boldsymbol{x}[j]\right)\right\}\right]\right| \\ &\le k \max_{1\le i\le k}\left\|\boldsymbol{w}_i^{(t)}\right\| \max_{1\le j\le n}\|\boldsymbol{x}[j]\| \\ &\le k\eta_1(t+1) \\ &\le \frac{1}{2}\end{aligned}$$

where the last inequality follows by the assumption on $\eta_1$. □

Lemma 5.3 follows by the following lemma.

**Lemma C.3.** *With probability at least $1 - 4e^{-\frac{m}{36}}$, for all $0 \le t \le T_1$ and all $i \in \mathcal{W}_0^+ \cap \mathcal{A}^+$ the following holds:*

1. $\boldsymbol{o}_1 \cdot \boldsymbol{w}_i^{(t)} \ge \frac{t\eta_1}{9}$.
2. *For all $j \ne 1$, it holds that $\boldsymbol{o}_j \cdot \boldsymbol{w}_i^{(t)} \le r$.*

*Proof.* We will prove the claim for $i \in \mathcal{W}_0^+ \cap \mathcal{A}^+$. We prove the two claims by induction on $t$. In the proof by induction we also show a third claim that: for all $\boldsymbol{x}_+ \in S_1^+$, $\boldsymbol{p}_t^{(i)}(\boldsymbol{x}_+) = \boldsymbol{o}_1$.

For the proof, we condition on the event:

$$\frac{|S_1^+|}{m_1}, \frac{|S_1^-|}{m_1} \ge \frac{m_1}{3} \tag{2}$$

This holds with probability at least $1 - 4e^{-\frac{m}{36}}$ by applying Hoeffding's inequality and a union bound (over positive and negative samples).

For $t = 0$, we have by definition for all $i \in \mathcal{W}_0^+ \cap \mathcal{A}^+$, $\boldsymbol{o}_1 \cdot \boldsymbol{w}_i^{(t)} > 0$. The second claim holds by the definition of the initialization. The third claim follows by the definition of $\mathcal{W}_0^+ \cap \mathcal{A}^+$.

Assume the three claims above hold for $t = T$. We will prove them for $t = T + 1$.

<u>Proof of Claim 1.</u> By the gradient update in the first layer, the following holds for $i \in \mathcal{W}_0^+ \cap \mathcal{A}^+$:

$$\begin{aligned}\boldsymbol{w}_i^{(T+1)} = \boldsymbol{w}_i^{(T)} &- \frac{\eta_1}{m_1}\sum_{\boldsymbol{x}_+ \in S_1^+} \ell'\left(N_{\text{CNN}}\left(\boldsymbol{x}_+; \left(W^{(T)}, a^{(0)}\right)\right)\right)\boldsymbol{p}_T^{(i)}(\boldsymbol{x}_+) \\ &+ \frac{\eta_1}{m_1}\sum_{\boldsymbol{x}_- \in S_1^-} \ell'\left(-N_{\text{CNN}}\left(\boldsymbol{x}_-; \left(W^{(T)}, a^{(0)}\right)\right)\right)\boldsymbol{p}_T^{(i)}(\boldsymbol{x}_-)\end{aligned} \tag{3}$$

where $l'(z) = -\frac{1}{1+e^z}$ is the derivative of the logistic loss. Note that for all $z$, $|\ell'(z)| \le 1$. Therefore, for all $\boldsymbol{x}_- \in S_1^-$, we have:

$$\left|\ell'\left(-N_{\text{CNN}}\left(\boldsymbol{x}_-; \left(W^{(T)}, a^{(0)}\right)\right)\right)\right| \le 1 \tag{4}$$

By Lemma C.2 we have for all $\boldsymbol{x} \in S_1$ $\left|N_{\text{CNN}}\left((; (W^{(t)}, a^{(0)}))\,\boldsymbol{x}\right)\right| \le \frac{1}{2}$. Therefore, for all $\boldsymbol{x}_+ \in S_1^+$:

$$\left|\ell'\left(N_{\text{CNN}}\left(\boldsymbol{x}_+; \left(W^{(T)}, a^{(0)}\right)\right)\right)\right| \ge \frac{1}{1+\sqrt{e}} \ge \frac{1}{3} \tag{5}$$

By the induction hypothesis, we have for $i \in \mathcal{W}_0^+ \cap \mathcal{A}^+$ and all $\boldsymbol{x}_+ \in S_1^+$ that $\boldsymbol{p}_T^{(i)}(\boldsymbol{x}_+) = \boldsymbol{o}_1$. Therefore we have:

$$\boldsymbol{p}_T^{(i)}(\boldsymbol{x}_+) \cdot \boldsymbol{o}_1 = 1 \tag{6}$$

For all $\boldsymbol{x}_- \in S_1^-$, we have $\boldsymbol{p}_T^{(i)}(\boldsymbol{x}_-) = \boldsymbol{o}_j$ for $j \neq 1$ that depends on $\boldsymbol{x}_-$. Therefore:

$$\boldsymbol{p}_T^{(i)}(\boldsymbol{x}_-) \cdot \boldsymbol{o}_1 = 0 \tag{7}$$

By the facts above we complete the proof of the first claim:

$$\begin{aligned}
\boldsymbol{w}_i^{(T+1)} \cdot \boldsymbol{o}_1 \underset{\text{Eq. 3,4,5}}{\geq} & \boldsymbol{w}_i^{(T)} \cdot \boldsymbol{o}_1 + \frac{\eta_1}{3m_1} \sum_{\boldsymbol{x}_+ \in S_1^+} \boldsymbol{p}_T^{(i)}(\boldsymbol{x}_+) \cdot \boldsymbol{o}_1 \\
& - \frac{\eta_1}{m_1} \sum_{\boldsymbol{x}_- \in S_1^-} \boldsymbol{p}_T^{(i)}(\boldsymbol{x}_-) \cdot \boldsymbol{o}_1 \\
& \underset{\text{Eq. 2,6,7}}{\geq} \boldsymbol{w}_i^{(T)} \cdot \boldsymbol{o}_1 + \frac{\eta_1}{9} \\
& \geq \frac{(T+1)\eta_1}{9}
\end{aligned} \tag{8}$$

where the last inequality follows from the induction hypothesis.

Proof of Claim 2. Since for all $\boldsymbol{x}_+ \in S_1^+$, $\boldsymbol{p}_T^{(i)}(\boldsymbol{x}_+) = \boldsymbol{o}_1$ we have for all $1 \leq j \leq d$, $j \neq 1$:

$$\boldsymbol{p}_T^{(i)}(\boldsymbol{x}_+) \cdot \boldsymbol{o}_j = 0 \tag{9}$$

By the facts (1) for all $\boldsymbol{x}_- \in S_1^-$ and $j \neq 1$ it holds that $\boldsymbol{p}_T^{(i)}(\boldsymbol{x}_-) \cdot \boldsymbol{o}_j \geq 0$ and (2) $l'(z) < 0$ for all $z$, we have:

$$\frac{\eta_1}{m_1} \sum_{\boldsymbol{x}_- \in S_1^-} \ell'\left(-N_{\text{CNN}}\left((; (W^{(T)}, a^{(0)})\right) \boldsymbol{x}_-)\right) \boldsymbol{p}_T^{(i)}(\boldsymbol{x}) \cdot \boldsymbol{o}_j \leq 0 \tag{10}$$

Therefore we have for $j \neq 1$:

$$\boldsymbol{w}_i^{(T+1)} \cdot \boldsymbol{o}_j \underset{\text{Eq.9,10}}{\leq} \boldsymbol{w}_i^{(T)} \cdot \boldsymbol{o}_j \leq r \tag{11}$$

where the right inequality follows by the induction hypothesis.

Proof of Claim 3. Since $r < \frac{\eta_1(T+1)}{9}$ we conclude by Eq. 8 and Eq. 11 that for all $\boldsymbol{x}_+ \in S_1^+$, $\boldsymbol{p}_{T+1}^{(i)}(\boldsymbol{x}_+) = \boldsymbol{o}_1$. $\square$

## D PROOF OF LEMMA 5.5

By Lemma C.1, for all $1 \leq t \leq T_1$ and $1 \leq i \leq k$, $\left\|\boldsymbol{w}_i^{(t)}\right\| \leq \eta_1(t+1)$. Therefore, for all $1 \leq j \leq d$ and $\boldsymbol{x}$ sampled from $\mathcal{D}$, $\boldsymbol{x}[j] \cdot \boldsymbol{w}_i^{(t)} \leq 2\eta_1 t$.

## E PROOF OF PART 3 OF THEOREM 5.1

Here we condition on the events of previous lemmas which hold with probability at least $1 - 4e^{-d} - 4e^{-\frac{m}{36}}$. For each $\boldsymbol{x}$ sampled from $\mathcal{D}$, define $\boldsymbol{z}(\boldsymbol{x}) \in \mathbb{R}^k$ such that for all $1 \leq i \leq k$, its $i$th entry is $z_i(\boldsymbol{x}) = \max_j \left\{\sigma\left(\boldsymbol{w}_i^{(T_1)} \cdot \boldsymbol{x}[j]\right)\right\}$. Notice that by Eq. 3 we have $z_i(\boldsymbol{x}) = \boldsymbol{w}_i^{(T_1)} \cdot \boldsymbol{p}_i^{(T_1)}(\boldsymbol{x})$. Define a new distribution of points $\mathcal{D}_{\boldsymbol{z}}$ over $\mathbb{R}^k \times \{\pm 1\}$, which samples a point $(\boldsymbol{z}(\boldsymbol{x}), y)$ where $(\boldsymbol{x}, y) \sim \mathcal{D}$.

Our goal is to show that $\mathcal{D}_{\boldsymbol{z}}$ is linearly separable and can be separated with a classifier of relatively low norm. Then, we will use recent results on logistic regression, which show that GD converges to low norm solutions. Therefore, by optimizing the second layer, $\text{LW}_{\text{CNN}}$ will converge to a low norm solution. Finally, we will apply norm-based generalization bounds to obtain a generalization guarantee for $\text{LW}_{\text{CNN}}$.

First we will show that $\mathcal{D}_{\boldsymbol{z}}$ is linearly separable. Indeed define $\boldsymbol{v}^* \in \mathbb{R}^k$ as follows. For $i \in \mathcal{W}_0^+ \cap \mathcal{A}^+$ let $\boldsymbol{v}_i^* = \frac{80d}{k\eta_1 T_1}$ and for $i \in \mathcal{W}_0^- \cap \mathcal{A}^-$ let $\boldsymbol{v}_i^* = -\frac{80d}{k\eta_1 T_1}$. Set all other entries of $\boldsymbol{v}^*$ to 0. Then for any $\boldsymbol{z}(\boldsymbol{x}_+)$ such that $(\boldsymbol{x}_+, 1) \sim \mathcal{D}$, we have:

$$\begin{aligned}
\boldsymbol{z}(\boldsymbol{x}_+) \cdot \boldsymbol{v}^* &= \frac{80d}{k\eta_1 T_1} \sum_{i \in \mathcal{W}_0^+ \cap \mathcal{A}^+} \boldsymbol{w}_i^{(T_1)} \cdot \boldsymbol{p}_i^{(T_1)}(\boldsymbol{x}_+) \\
&- \frac{80d}{k\eta_1 T_1} \sum_{i \in \mathcal{W}_0^- \cap \mathcal{A}^-} \boldsymbol{w}_i^{(T_1)} \cdot \boldsymbol{p}_i^{(T_1)}(\boldsymbol{x}_+) \\
&> \left(\frac{80d}{k\eta_1 T_1}\right)\left(\frac{k}{4d}\right)\left(\frac{\eta_1 T_1}{10}\right) \\
&- \left(\frac{80d}{k\eta_1 T_1}\right)\left(\frac{k}{d}\right)\left(\frac{\eta_1 T_1}{80}\right) \\
&= 1
\end{aligned}$$

where the inequality follows by Lemma 5.2, Lemma 5.3 and Corollary 5.4. By symmetry, we have $-\boldsymbol{z}(\boldsymbol{x}_-) \cdot \boldsymbol{v}^* > 1$ for all $(\boldsymbol{x}_-, -1) \sim \mathcal{D}$.

Next, we proceed to apply Theorem 3 in Soudry et al. [2018]. It requires that $\eta_2 < 2\beta^{-1}\sigma_{\max}^{-2}(Z)\, m_2$,[2] where $\beta$ is the smoothness parameter of the logistic loss, $Z \in \mathbb{R}^{k \times m_2}$ is the matrix which contains $\boldsymbol{z}(\boldsymbol{x}_{i+\lceil \frac{m}{2} \rceil})$ in its $i$th column and $\sigma_{\max}(Z)$ is the maximum singular value of $Z$. In our setting, $\beta = 1$ and by Lemma 5.5 $\sigma_{\max}^2(Z) \le \|Z\|_F^2 \le 4m_2 k\eta_1^2 T_1^2 \le \frac{m_2}{4k}$. Thus, by our assumption $\eta_2 < 8k \le 2\sigma_{\max}^{-2}(Z)\, m_2$ holds.

Therefore, by this theorem we are guaranteed that:

$$\lim_{t \to \infty} \frac{\boldsymbol{a}^{(t)}}{\left\|\boldsymbol{a}^{(t)}\right\|} = \frac{\hat{\boldsymbol{a}}}{\|\hat{\boldsymbol{a}}\|} \tag{12}$$

where

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{v} \in \mathbb{R}^k} \|\boldsymbol{v}\|^2 \;\text{ s.t. }\; \forall i \quad y_i \boldsymbol{v} \cdot \boldsymbol{z}(\boldsymbol{x}_i) \ge 1 \tag{13}$$

Specifically, gradient descent converges to zero training loss, i.e., $\lim_{T_2 \to \infty} \mathcal{L}_2\left((W_{T_1}, \boldsymbol{a}_{T_2})\right) = 0$.

By optimality of $\hat{\boldsymbol{a}}$ and Lemma 5.2 we have $\|\hat{\boldsymbol{a}}\|^2 \le \|\boldsymbol{v}^*\|^2 \le \frac{80^2 d^2}{k^2 \eta_1^2 T_1^2} \frac{2k}{d} = \frac{2 \cdot 80^2 d}{k\eta_1^2 T_1^2}$. Furthermore, $\|\boldsymbol{z}(\boldsymbol{x})\|^2 \le 4k\eta_1^2 T_1^2$ by Lemma 5.5. Therefore, we have $\|\hat{\boldsymbol{a}}\|^2 \|\boldsymbol{z}(\boldsymbol{x})\|^2 = O(d)$. Thus, by a standard margin generalization bound (e.g. Theorem 26.13 in Shalev-Shwartz and Ben-David [2014] or Bartlett and Mendelson [2002]) we have with probability at least $1 - \delta$:

$$\begin{aligned}
&\lim_{T_2 \to \infty} \mathbb{P}_{(\boldsymbol{x}, y) \sim \mathcal{D}}\left(\text{sign}\left(N_{\text{CNN}}\left(\boldsymbol{x}; \left(W^{(T_1)}, \boldsymbol{a}^{(T_2)}\right)\right)\right) \neq y\right) \\
&= \mathbb{P}_{(\boldsymbol{x}, y) \sim \mathcal{D}}\left(\text{sign}\left(N_{\text{CNN}}\left(\boldsymbol{x}; \left(W^{(T_1)}, \frac{\hat{\boldsymbol{a}}}{\|\hat{\boldsymbol{a}}\|}\right)\right)\right) \neq y\right) \\
&= O\left(\sqrt{\frac{d}{m}}\right)
\end{aligned}$$

where $O$ hides an additive term which depends on $\delta$.

## References


Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Ziwei Ji and Matus Telgarsky. A refined primal-dual analysis of the implicit bias. *arXiv preprint arXiv:1906.04540*, 2019.


[2]We added the factor $m_2$ because Soudry et al. [2018] consider the empirical loss without dividing by the number of samples.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.