Foundations of Deep Learning – Exercise 4

Nadav Gat, Yoav Nagel

IDs: 207291683, 205947336

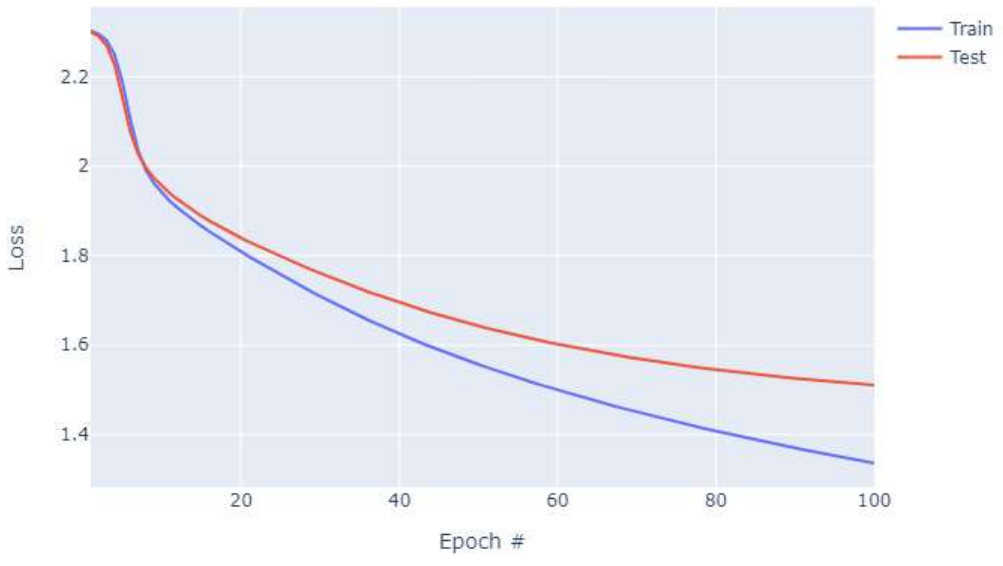
Part 1

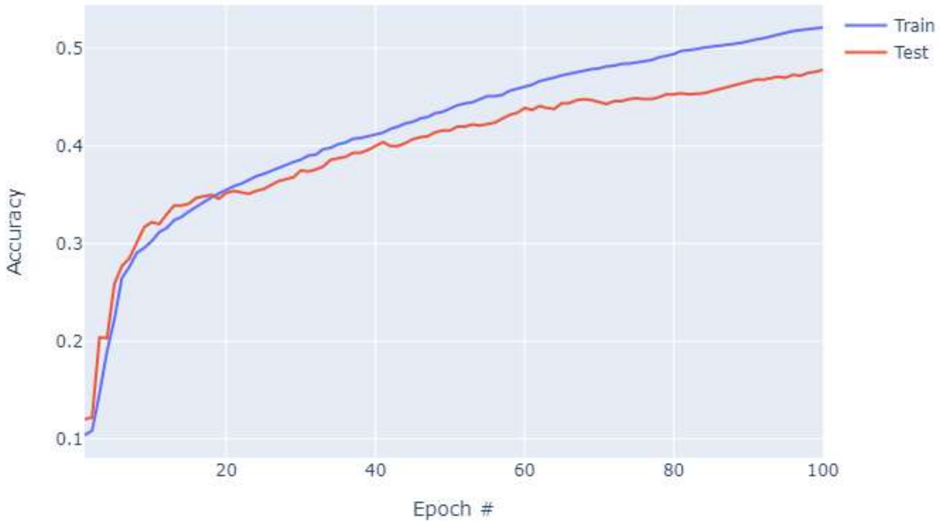
In all experiments we used 10% of CIFAR10 (i.e. 5,000 samples of train set and 1,000 of test set) and a NN with 2 convolutional layers (with max-pooling between them), followed by 3 fully-connected layers. The model has ~60,000 parameters, over 12 times more than the train set size. We used the ADAM optimizer, with different learning rates and number of epochs for each experiment.

Experiment (i):

We used a learning rate of 0.00005 and ran for 100 epochs.

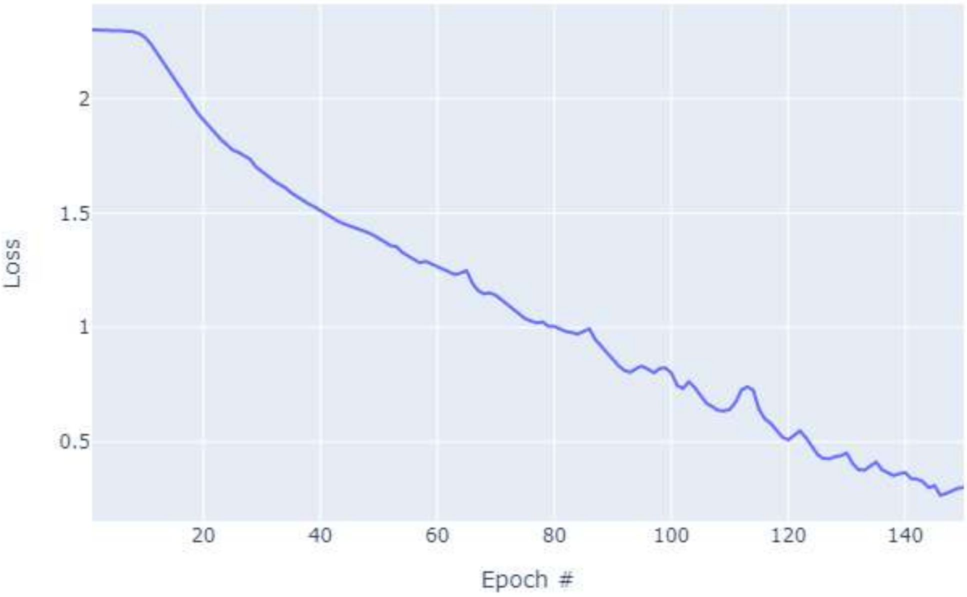
From the graph below, where train and test loss are plotted over time, it is clear that the model generalizes very well, a gap between train and test loss starts at around epoch 30, but it is relatively small and the test loss for the model continues to improve significantly together with train loss. The same thing is true for train and test accuracy.

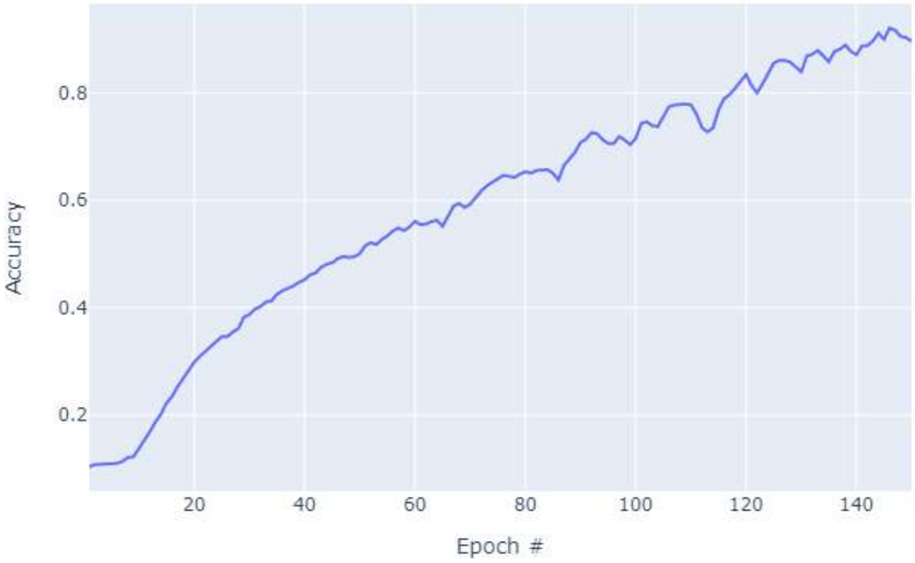




Experiment (ii):

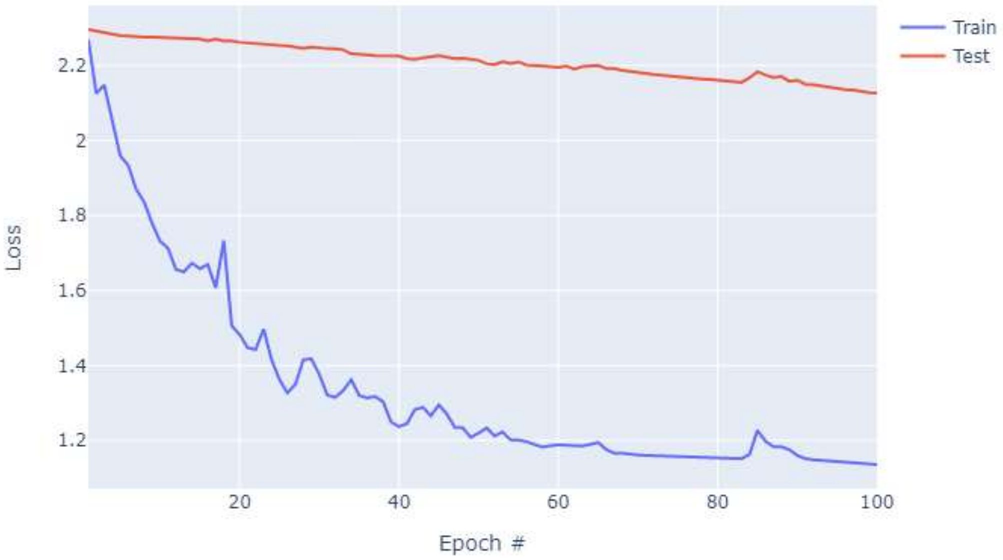
We generated 5,000 samples of random data and labels, and used a learning rate of 0.001 for quicker convergence, and ran for 150 epochs. The graph shows that the loss decreases monotonically for most epochs and converges to near 0, and accuracy increases monotonically, reaching more than 98% by the last epoch.

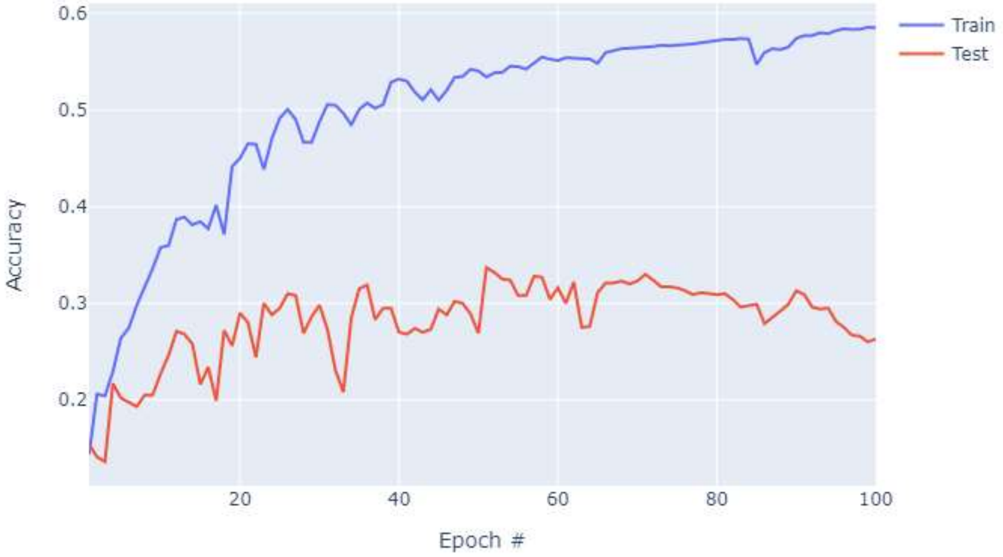




Experiment (iii):

We trained with 5,000 samples from CIFAR10 and 5,000 samples generated randomly. We used a learning rate of 0.00033 and ran for 100 epochs, and computed test loss over 1,000 samples from CIFAR10. From the graph below of train and test loss it is clear that this model generalizes much worse than the one in (i), but there is still distinct improvement in test loss, meaning the model performs better than trivially. With more hyperparameter tuning and many more epochs, it is reasonable to assume the model will continue to improve and reach a low test loss as well. The model's improvement is easier to notice in the accuracy graph – it achieves ~30% test accuracy by the end, which is far better than the trivial 10% that can be expected for a classification problem with 10 classes.

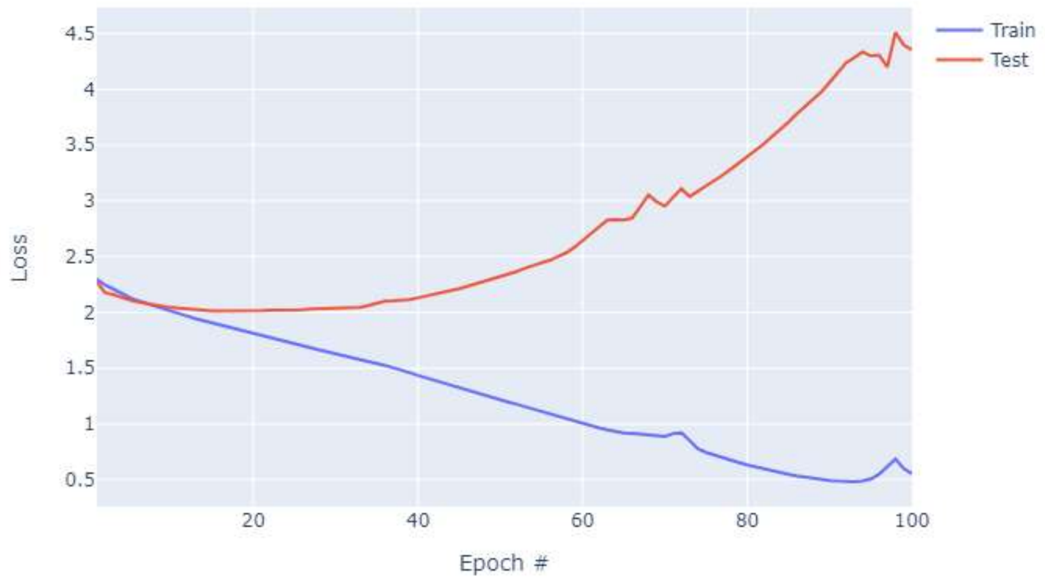


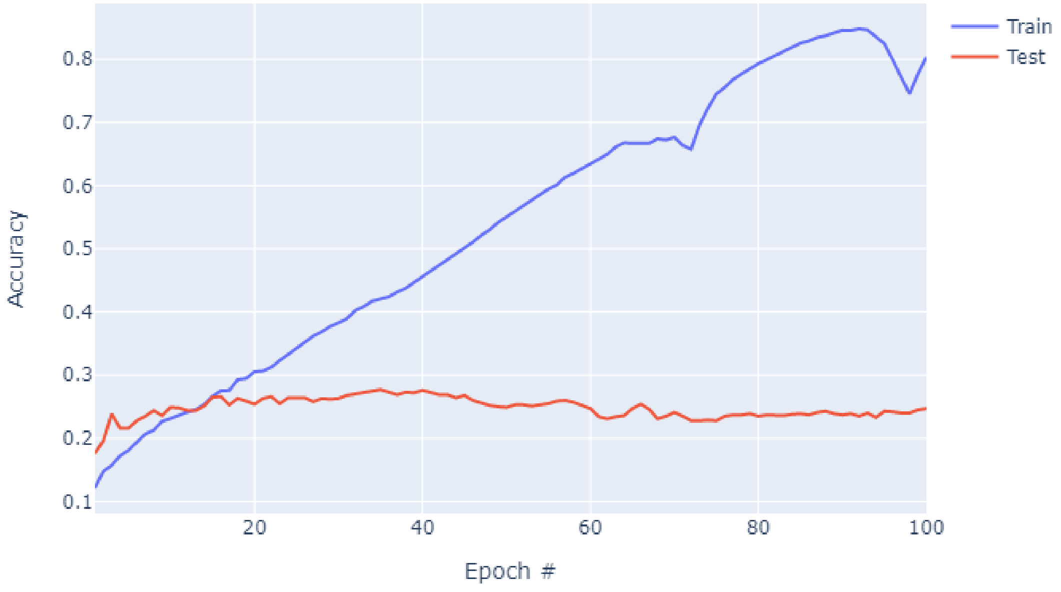


Experiment iv:

In this experiment we took 2,500 samples from CIFAR10 as-is, and another 2,500 where we added 5 to the label (modulus 10) as adversarial samples. We trained with a learning rate of 0.0003 for 100 epochs.

As can be seen in the graph below, the train loss converged to near zero, while the test loss went down for ~20 epochs, and then increased significantly until the end of the run, with an increasing slope. A similar phenomena can be seen in the accuracy graph – train accuracy keeps improving and reaches ~80%, while test accuracy improves slightly but gets stuck at ~20% (about 10% less than in the previous experiment).





Part 2

Ex. 1a

We'll first calculate the size of $\mathcal{H}_r$. It has $2Ndr$ parameters, and each can 2^b values (since that is the number of bits needed for each value). The number of hypotheses is bounded from above by the number of different assignments to the parameters (different assignments may lead to the same end-to-end hypothesis so this is not a tight bound). Therefore, we have:

$$|\mathcal{H}_r| \leq 2Ndr2^b = Ndr2^{b+1} \leq \exp_2\big((b+1)\cdot\lceil\log_2(Ndr)\rceil\big) \coloneqq 2^\beta$$

Where $\exp_2(x) = 2^x$, $\lceil x\rceil$ is the closest integer to x, from above.

Let $\hat{h} \in \mathcal{H}$ be a hypothesis. From the proposition we saw in class we have that for any $\delta \in (0,1)$ w.p. $\geq 1-\delta$ over $S \sim D^m$:

$$L_D(\hat{h}) - L_S(\hat{h}) \leq \sqrt{\frac{(\beta+1)\ln 2 + \ln\left(\frac{1}{\delta}\right)}{2m}} + 2\rho\cdot d(\hat{h},\mathcal{H}_r)$$

We'll now calculate the last term. Let $W_1, \dots, W_n$ be weight matrices that yield $\hat{h}$. Denote by $W_1', \dots, W_N'$ their closest rank r approximations, and by h' the resulting hypothesis. Recall that the distance is defined:

$$\begin{aligned} d(\hat{h},\mathcal{H}_r) \coloneqq \min_{h\in\mathcal{H}_r}\sup_{x\in\mathcal{X}}\big\|\hat{h}(x) - h(x)\big\| &\leq \sup_{x\in\mathcal{X}}\big\|\hat{h}(x) - h'(x)\big\| \\ &= \sup_{x:\|x\|\leq 1}\|W_N\sigma(\dots\sigma(W_1x)\dots) - W_N'\sigma(\dots\sigma(W_1'x)\dots)\| \end{aligned}$$

For $n \in [N]$, denote:

$$e_n(x) \coloneqq \|W_n\sigma(\dots\sigma(W_1x)\dots) - W_n'\sigma(\dots\sigma(W_1'x)\dots)\|$$

Then:

$$e_1(x) = \|W_1x - W_1'x\| = \|(W_1 - W_1')\cdot x\| \leq \|W_1 - W_1'\|_s \cdot \|x\|$$

Where $\|W\|_s$ is the spectral norm of W, and we used the multiplicativeness of norms. For $n > 1$:

$$
\begin{aligned}
e_n(x) = {} & \big\|W_n\sigma\big(W_{n-1}\sigma(\ldots\sigma(W_1x)\ldots)\big) - W_n'\sigma\big(W_{n-1}\sigma(\ldots\sigma(W_1x)\ldots)\big) \\
& + W_n'\sigma\big(W_{n-1}\sigma(\ldots\sigma(W_1x)\ldots)\big) - W_n'\sigma\big(W_{n-1}'\sigma(\ldots\sigma(W_1x)\ldots)\big)\big\| \\
& \leq \big\|W_n\sigma\big(W_{n-1}\sigma(\ldots\sigma(W_1x)\ldots)\big) - W_n'\sigma\big(W_{n-1}\sigma(\ldots\sigma(W_1x)\ldots)\big)\big\| \\
& + \big\|W_n'\sigma\big(W_{n-1}\sigma(\ldots\sigma(W_1x)\ldots)\big) - W_n'\sigma\big(W_{n-1}'\sigma(\ldots\sigma(W_1x)\ldots)\big)\big\| \\
& = \big\|(W_n - W_n')\sigma\big(W_{n-1}\sigma(\ldots\sigma(W_1x)\ldots)\big)\big\| \\
& + \Big\|W_n'\left(\sigma\big(W_{n-1}\sigma(\ldots\sigma(W_1x)\ldots)\big) - \sigma\big(W_{n-1}'\sigma(\ldots\sigma(W_1x)\ldots)\big)\right)\Big\| \\
& \leq \|W_n - W_n'\|_s \cdot \big\|\sigma\big(W_{n-1}\sigma(\ldots\sigma(W_1x)\ldots)\big)\big\| + \|W_n'\|_s \\
& \cdot \big\|\sigma\big(W_{n-1}\sigma(\ldots\sigma(W_1x)\ldots)\big) - \sigma\big(W_{n-1}'\sigma(\ldots\sigma(W_1x)\ldots)\big)\big\|
\end{aligned}
$$

We now use the fact that for any $v \in \mathbb{R}^d$: $\|\sigma(v)\| \leq \gamma\|v\|$ and the fact that the largest singular value of the best rank r approx. of W is equal to that of W. This is true since the best rank r approx. for $W_n = \sum_{i=1}^{d} \sigma_i u_i v_i^T$ where $|\sigma_1| \geq \cdots \geq |\sigma_d|$ is $\sum_{i=1}^{r} \sigma_i u_i v_i^T$, which has the same largest singular value. We then get:

$$
\begin{aligned}
e_n(x) \leq {} & \|W_n - W_n'\|_s \cdot \gamma \cdot \|W_{n-1}\sigma(\ldots\sigma(W_1x)\ldots)\| + \|W_n\|_s \cdot \gamma \\
& \cdot \big\|\big(W_{n-1}\sigma(\ldots\sigma(W_1x)\ldots)\big) - \big(W_{n-1}'\sigma(\ldots\sigma(W_1x)\ldots)\big)\big\| \\
& = \|W_n - W_n'\|_s \cdot \gamma \cdot \|W_{n-1}\sigma(\ldots\sigma(W_1x)\ldots)\| + \|W_n\|_s \cdot \gamma \cdot e_{n-1}(x) \leq \cdots \\
& \leq \|W_n - W_n'\|_s \cdot \gamma \cdot \|W_{n-1}\|_s \cdot \gamma \cdot \ldots \cdot \gamma\|W_1\|_s \cdot \|x\| + \|W_n\|_s \cdot \gamma \cdot e_{n-1}(x)
\end{aligned}
$$

Where the last inequality is attained by iteratively using $\|\sigma(v)\| \leq \gamma\|v\|$ and norm multiplicativeness. We then get:

$$
e_n(x) \leq \|W_n - W_n'\|_s \cdot \gamma^{n-1} \cdot \prod_{j=1}^{n-1} \|W_j\|_s + \|W_n\|_s \cdot \gamma \cdot e_{n-1}(x)
$$

And by induction we have:

$$
e_N(x) \leq \gamma^{N-1} \cdot \sum_{n=1}^{N} \prod_{j \in [N]\{n\}} \|W_j\|_s \cdot \|W_n - W_n'\|_s \cdot \|x\|
$$

Plugging back into the inequality we had for $d\big(\hat{h}, \mathcal{H}_r\big)$:

$$
d\big(\hat{h}, \mathcal{H}_r\big) \leq \sup_{x:\|x\|\leq 1} e_N(x) \leq \gamma^{N-1} \cdot \sum_{n=1}^{N} \prod_{j \in [N]\{n\}} \|W_j\|_s \cdot \|W_n - W_n'\|_s
$$

And finally we have:

$$
\begin{aligned}
&L_D(\hat{h}) - L_S(\hat{h}) \\
&\qquad \leq \sqrt{\frac{\left((b+1)\cdot\lceil\log_2(Ndr)\rceil + 1\right)\ln 2 + \ln\left(\frac{1}{\delta}\right)}{2m}} + 2\rho\cdot\gamma^{N-1} \\
&\qquad \cdot \sum_{n=1}^{N} \prod_{j\in[N]\{n\}} \left\|W_j\right\|_s \cdot \|W_n - W_n'\|_s
\end{aligned}
$$

Ex. 1b

Let $\hat{h}\in\mathcal{H}, \delta\in(0,1)$ and define $\delta' := \frac{\delta}{d}$. From (a) we have that w.p. $\geq 1-\delta'$ for any fixed $r\in[d]$:

$$
\begin{aligned}
&L_D(\hat{h}) - L_S(\hat{h}) \\
&\qquad \leq \sqrt{\frac{\left((b+1)\cdot\lceil\log_2(Ndr)\rceil + 1\right)\ln 2 + \ln\left(\frac{d}{\delta}\right)}{2m}} + 2\rho\cdot\gamma^{N-1} \\
&\qquad \cdot \sum_{n=1}^{N} \prod_{j\in[N]\{n\}} \left\|W_j\right\|_s \cdot \|W_n - W_n'\|_s
\end{aligned}
$$

Denoting the RHS as a_r, we can rewrite the previous conclusion as:

$$
\forall r\in[d].\, \mathbb{P}\left[L_D(\hat{h}) - L_S(\hat{h}) \geq a_r\right] \leq \delta' = \frac{\delta}{d}
$$

And using union bound we then get:

$$
\mathbb{P}\left[\exists r\in[d]\; s.t.\; L_D(\hat{h}) - L_S(\hat{h}) \geq a_r\right] \leq d\cdot\frac{\delta}{d} = \delta
$$

Meaning:

$$
\mathbb{P}\left[\forall r\in[d], L_D(\hat{h}) - L_S(\hat{h}) \leq a_r\right] \geq 1-\delta
$$

And this is the generalization bound for $\mathcal{H}$.

Foundation of Deep Leaning - Ex. 4

Nadav Gat & Yoav Nagel

IDs: 207291683 & 205947336

Part II

Generalization Bounds

Rademacher complexity and norms

Exercise 1. Let us define for any $k \in \mathbb{N}^+$ the hypothesis class

$$\mathcal{H}_k = \left\{ h_\theta : \mathcal{X} \to \mathcal{Y} \mid \theta \in \mathbb{R}^p, \frac{1}{k+1} \leq \|\theta\|_\infty \leq 0.5 \right\} \cup \left\{ h_\theta : \mathcal{X} \to \mathcal{Y} \mid \theta = 0 \in \mathbb{R}^p \right\}$$

One can easily notice that $\bigcup_{k=1}^{\infty} \mathcal{H}_k = \mathcal{H}$, and we can apply an inequality seen in class. By this inequality, w.p $\geq 1 - \frac{\delta}{2}$ we have $\forall k \in \mathbb{N}, h \in \mathcal{H}_k$ that

$$\mathcal{L}_D(h) - \mathcal{L}_S(h) \leq 2R(\ell \circ \mathcal{H}_k \circ S) + \sqrt{\frac{2\ln\left(\frac{2\pi^2}{3}k^2 \cdot \frac{2}{\delta}\right)}{m}} \cdot 4 \tag{*}$$

Since we assume $\mathbb{E}_S[R(\ell \circ \mathcal{H}_\Theta \circ S)] = Volume(\Theta)$ for any $\Theta \subseteq \{\theta \in \mathbb{R}^p \mid \|\theta\|_\infty \leq 0.5\}$ we have that $\forall k \in \mathbb{N}^+$

$$\mathbb{E}_S[R(\ell \circ \mathcal{H}_k \circ S)] = \frac{1}{2^p} - \frac{1}{(k+1)^p}$$

and by Markov's inequality, we can derive that w.p $\geq 1 - \frac{\delta}{2}$

$$R(\ell \circ \mathcal{H}_k \circ S) \leq \frac{2}{\delta} \cdot \mathbb{E}_S[R(\ell \circ \mathcal{H}_k \circ S)] \tag{**}$$

Since $A \subseteq B \Rightarrow P(A) \leq P(B)$ for any events A, B , we deduce

$$\begin{aligned}
&\mathbb{P}_S\left(\forall k \in \mathbb{N}, h \in \mathcal{H}_k : \mathcal{L}_D(h) - \mathcal{L}_S(h) \leq 2 \cdot \frac{2}{\delta}\mathbb{E}_S[R(\ell \circ \mathcal{H}_k \circ S)] + \sqrt{\frac{2\ln\left(\frac{2\pi^2}{3}k^2 \cdot \frac{2}{\delta}\right)}{m}} \cdot 4\right) \\
\geq &\mathbb{P}_S(\forall k \in \mathbb{N}, h \in \mathcal{H}_k : (*) \cap (**)) \\
\geq &\mathbb{P}_S(\forall k \in \mathbb{N}, h \in \mathcal{H}_k : (*)) + \mathbb{P}_S(\forall k \in \mathbb{N}, h \in \mathcal{H}_k : (**)) - 1 \\
\geq &\left(1 - \frac{\delta}{2}\right) + \left(1 - \frac{\delta}{2}\right) - 1 = 1 - \delta
\end{aligned}$$

So, by simplifying the above expression and plugging in the known expectation, we have our generalization bound:

w.p $\geq 1-\delta$ over $S \sim D^m$

$$\forall k \in \mathbb{N}^+, h \in \mathcal{H}_k : \mathcal{L}_D(h) - \mathcal{L}_S(h) \leq 4\left(\frac{1}{\delta}\left(\frac{1}{2^p} - \frac{1}{(k+1)^p}\right) + \sqrt{\frac{2\ln\left(\frac{2\pi^2}{3}k^2 \cdot \frac{2}{\delta}\right)}{m}}\right)$$

This inequality gives tighter bounds for lower k values, which guarantee higher max-norms, and therefore this suits our knowledge of the implicit regularization.

PAC-Bayes

Exercise 1. By definition, the density function of $\mathcal{N}(\mu,\Sigma)$ over $\mathbb{R}^n$ is given by

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

and KL divergence of distributions P, Q is defined as

$$KL(P \parallel Q) = \mathbb{E}_P\left[\ln\left(\frac{P(x)}{Q(x)}\right)\right]$$

Let us fix $P = \mathcal{N}(\mu_0,\Sigma_0), Q = \mathcal{N}(\mu_1,\Sigma_1)$. Developing the inner term will lead us to

$$\begin{aligned}
\ln\left(\frac{P(x)}{Q(x)}\right) &= \ln\left(\frac{\frac{1}{|\Sigma_0|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right)}{\frac{1}{|\Sigma_1|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right)}\right)\\
&= \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\\
&= \frac{1}{2}\left(\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right)
\end{aligned}$$

strictly by using logarithm rules and straight-forward algebra. Plugging it in KL divergence gives us

$$\begin{aligned}
KL(P \parallel Q) &= \mathbb{E}_P\left[\frac{1}{2}\left(\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right)\right]\\
&= \frac{1}{2}\left(\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + \mathbb{E}_P\left[\operatorname{tr}\left((x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right)\right] - \mathbb{E}_P\left[\operatorname{tr}\left((x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right)\right]\right) && \text{by linearity of } \mathbb{E} \text{ and by taking trace on real values}\\
&= \frac{1}{2}\left(\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + \mathbb{E}_P\left[\operatorname{tr}\left(\Sigma_1^{-1}(x-\mu_1)(x-\mu_1)^T\right)\right] - \mathbb{E}_P\left[\operatorname{tr}\left(\Sigma_0^{-1}(x-\mu_0)(x-\mu_0)^T\right)\right]\right) && \text{trace is cyclic}\\
&= \frac{1}{2}\left(\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + \operatorname{tr}\left(\mathbb{E}_P\left[\Sigma_1^{-1}(x-\mu_1)(x-\mu_1)^T\right]\right) - \operatorname{tr}\left(\mathbb{E}_P\left[\Sigma_0^{-1}(x-\mu_0)(x-\mu_0)^T\right]\right)\right) && \text{linearity of } \mathbb{E}\\
&= \frac{1}{2}\left(\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + \operatorname{tr}\left(\Sigma_1^{-1}\mathbb{E}_P\left[xx^T - \mu_1x^T - x\mu_1^T + \mu_1\mu_1^T\right]\right) - \operatorname{tr}\left(\Sigma_0^{-1}\mathbb{E}_P\left[(x-\mu_0)(x-\mu_0)^T\right]\right)\right)
\end{aligned}$$

Now, we can use the definition of the covariance matrix $\Sigma_0 = \mathbb{E}_P\left[(x-\mu_0)(x-\mu_0)^T\right]$, from which we can also take $\mathbb{E}_P\left[xx^T\right] = \Sigma_0 + \mu_0\mu_0^T$, also using $\mathbb{E}_P[x] = \mu_0$. Plugging all of this in

$$\begin{aligned}
KL(P \parallel Q) &= \frac{1}{2}\left(\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + \operatorname{tr}\left(\Sigma_1^{-1}\left(\Sigma_0 + \mu_0\mu_0^T - \mu_1\mu_0^T - \mu_0\mu_1^T + \mu_1\mu_1^T\right)\right) - \operatorname{tr}(I_n)\right) \\
&= \frac{1}{2}\left(\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + \operatorname{tr}\left(\Sigma_1^{-1}\Sigma_0\right) + \operatorname{tr}\left(\Sigma_1^{-1}(\mu_1-\mu_0)(\mu_1-\mu_0)^T\right) - n\right) && \text{linearity of trace} \\
&= \frac{1}{2}\left(\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + \operatorname{tr}\left(\Sigma_1^{-1}\Sigma_0\right) + \operatorname{tr}\left((\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0)\right) - n\right) && \text{trace is cyclic} \\
&= \frac{1}{2}\left(\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + \operatorname{tr}\left(\Sigma_1^{-1}\Sigma_0\right) + (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0) - n\right) && \text{trace of real value}
\end{aligned}$$

which is the desired expression.

Exercise 2. Let us have this set of parameters $\{\theta_1, ..., \theta_k\}$. We shall assume prior distributions P_i for any $1 \leq i \leq k$ defined to be $\mathcal{N}\left(\theta_i, \sigma^2 \cdot I\right)$ over $\mathbb{R}^n$. We shall define Q as a distribution over the hypothesis space, as $\mathcal{N}\left(\hat{\theta}, \bar{\sigma}^2 \cdot I\right)$ where $\hat{\theta}$ are the parameters returned by training algorithm and $\bar{\sigma}^2$ was fixed in advance. So, as seen in class and by the previous section

$$KL(Q \parallel P_i) = \frac{1}{2}\left(n \cdot \frac{1}{\sigma^2}\bar{\sigma}^2 + \frac{1}{\sigma^2}\|\hat{\theta} - \theta_i\|^2 - n + n \cdot \ln\left(\sigma^2\right) - n \cdot \ln\left(\bar{\sigma}^2\right)\right)$$

and in class we also saw that by the choice of $\bar{\sigma} = \sigma$ which we will make, we minimize the KL divergence, getting

$$KL(Q \parallel P_i) = \frac{1}{2\sigma^2}\|\hat{\theta} - \theta_i\|^2$$

We shall use the theorem seen in class, and for a fixed i we can deduce that w.p $\geq 1 - \frac{\delta}{k}$

$$L_D(Q) \leq L_S(Q) + \sqrt{\frac{KL(Q \parallel P_i) + \ln\left(\frac{2m}{\delta}k\right)}{2(m-1)}} = L_S(Q) + \sqrt{\frac{\frac{1}{2\sigma^2}\|\hat{\theta} - \theta_i\|^2 + \ln\left(\frac{2m}{\delta}k\right)}{2(m-1)}}$$

Taking union bound gives us w.p $\geq 1 - \delta$ that $\forall i \in [k]$ the expression above holds. Equivalently,

$$L_D(Q) \leq L_S(Q) + \min_{1 \leq i \leq k}\sqrt{\frac{\frac{1}{2\sigma^2}\|\hat{\theta} - \theta_i\|^2 + \ln\left(\frac{2m}{\delta}k\right)}{2(m-1)}} = L_S(Q) + \sqrt{\frac{\frac{1}{2\sigma^2}\min_{1 \leq i \leq k}\|\hat{\theta} - \theta_i\|^2 + \ln\left(\frac{2m}{\delta}k\right)}{2(m-1)}}$$

Indeed, knowing the implicit regularization tends to flat minima, we obtain solutions that bring lower $L_S(Q)$. The tendency to bring the solution close to a certain θ_i will result in lower values of the second expression $\sqrt{\frac{\frac{1}{2\sigma^2}\min_{1 \leq i \leq k}\|\hat{\theta} - \theta_i\|^2 + \ln\left(\frac{2m}{\delta}k\right)}{2(m-1)}}$ since it tends to minimize $\|\hat{\theta} - \theta_i\|^2$. Thus, this implicit regularization achieves lower upper bound for the population loss, accounting for this bound we showed.

Part III

Implicit Regularization

Linear Regression

Exercise. The proof starts similarly to what we have already seen in class. $\forall i \in [m]$ and $w \in \mathbb{R}^d$:

$$\nabla \ell_{(x_i,y_i)}(w) = (\langle x_i, w\rangle - y_i)\, x_i$$

and this implies that for any $t \in \mathbb{N} \cup \{0\}$ $w^{(t+1)} - w^{(t)} \in \operatorname{span}\{x_i\}_{i=1}^m$. Since we assume $w^{(0)} = a$, $\forall t \in \mathbb{N}$ we have that $w^{(t)}$ lies in an affine space $a + \operatorname{span}\{x_i\}_{i=1}^m$, meaning we have some $\{\alpha_i\}_{i=1}^m \subseteq \mathbb{R}$ for which $w^{(t)} = a + \sum_{i=1}^m \alpha_i x_i$. This affine space is topologically closed, and therefore $w^{(\infty)} := \lim\limits_{t\to\infty} w^{(t)}$ must lie in this affine space, i.e

$$\exists \bar{r} \in \mathbb{R}^m : w^{(\infty)} = a + X\bar{r}$$

where $X := [x_1, x_2, ..., x_m] \in \mathbb{R}^{d,m}$, and assuming $\operatorname{rank} X = m$ (as the setting shown in class). Since $L_S\left(w^{(\infty)}\right) = 0$:

$$\begin{aligned} X^T w^{(\infty)} = y &\Rightarrow X^T(a + X\bar{r}) = y \\ &\Rightarrow X^T X\bar{r} = y - X^T a \\ &\Rightarrow \bar{r} = \left(X^T X\right)^{-1}\left(y - X^T a\right) \end{aligned}$$

And therefore

$$w^{(\infty)} = a + X\left(X^T X\right)^{-1}\left(y - X^T a\right)$$

Taking the Euclidean norm over the solution, we can show an upper bound using triangle inequality

$$\begin{aligned} \|w^{(\infty)}\| &= \|a + X\left(X^T X\right)^{-1}\left(y - X^T a\right)\| \\ &= \|a + X\left(X^T X\right)^{-1} y - X\left(X^T X\right)^{-1} X^T a\| \\ &\le \|a - X\left(X^T X\right)^{-1} X^T a\| + \|X\left(X^T X\right)^{-1} y\| \end{aligned}$$

Since we know from the lecture that $\hat{w} = X\left(X^T X\right)^{-1} y$ is the solution that gives us the min norm, we know that the distance of the solution we achieved in this particular setting from min norm is at most $\|a - X\left(X^T X\right)^{-1} X^T a\|$. We'll claim that $P_\perp a = a - X\left(X^T X\right)^{-1} X^T a$ and that would conclude our proof.

Let us assume we have some vector on $\operatorname{span}\{x_i\}_{i=1}^m$ that is the closest to a, so it could be written as Xb for some $b \in \mathbb{R}^m$ which gives $P_\perp a = a - Xb$. Therefore, $a - Xb$ is perpendicular to $\operatorname{span}\{x_i\}_{i=1}^m$, and this implies

$$X^T(a - Xb) = 0 \Rightarrow Xb = X\left(X^T X\right)^{-1} X^T a$$

which leads us to deduce that the projection is the desired expression, and that $\|w^{(\infty)}\| - \|\hat{w}\| \le \|P_\perp a\|$.