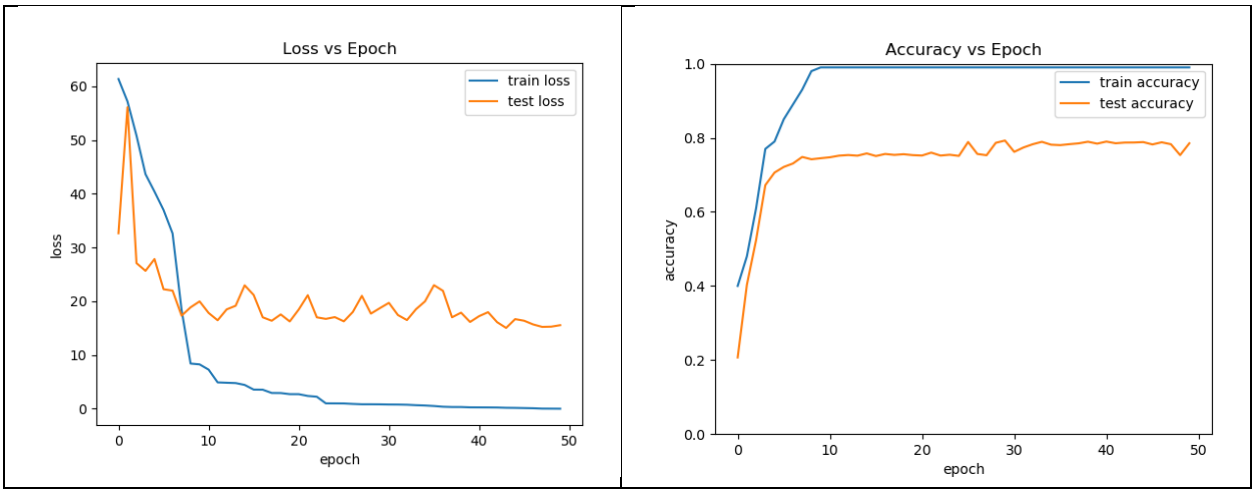
Assignment 4: Generalization

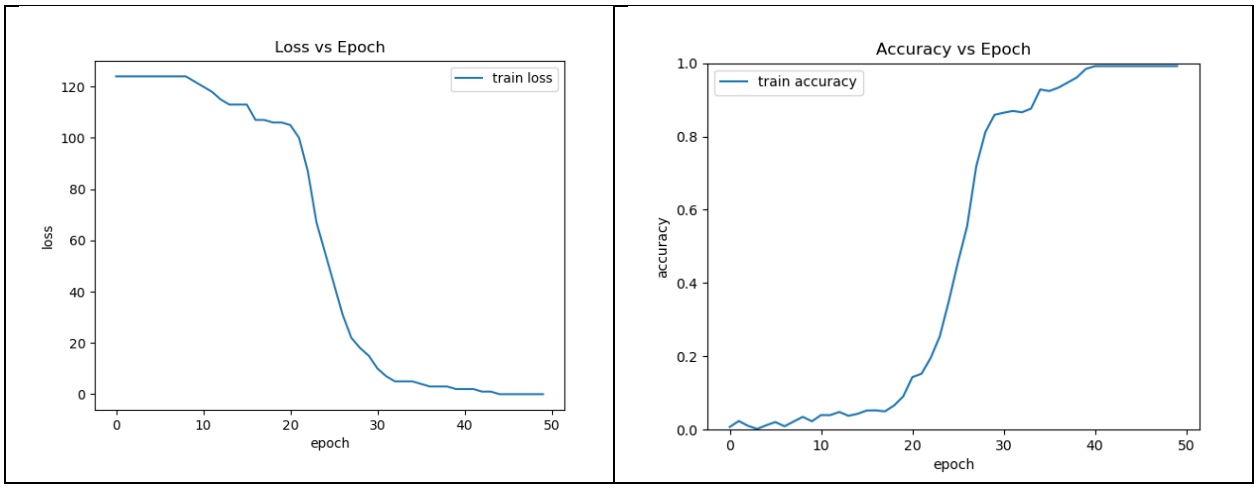
Dana Arad 311183321
Amit Elyasi 316291434

Part 1: Empirical Phenomena

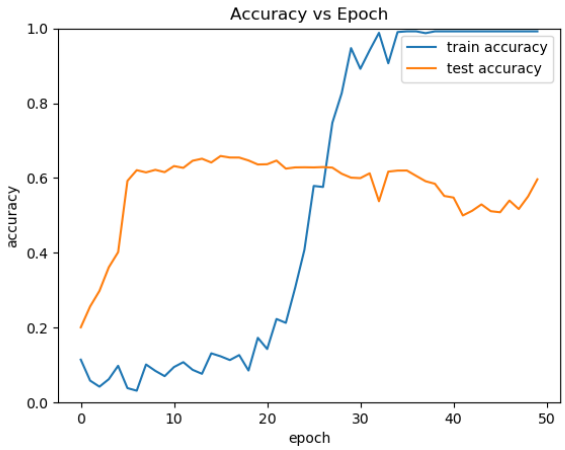
1. Generalization without any regularization (wd, dropout)



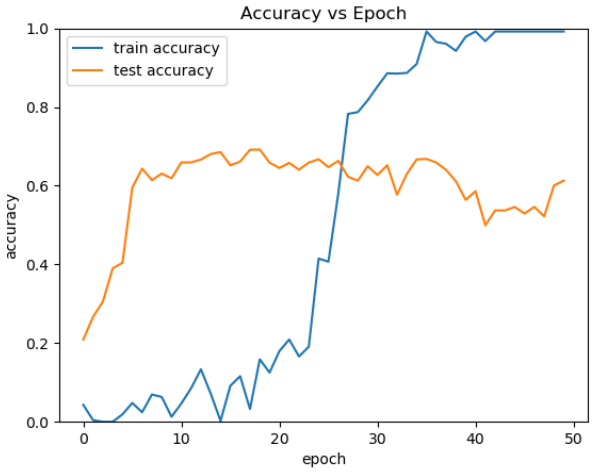
2. Zero training error on random labels



3. Better-then-random test accuracy when training on random labels



4. Adversarially changing the labels increases the test accuracy:



Part 2 : Generalization Bounds

Compression

1.

a) Fixing $r \in [d]$ and deriving a generalization bound for $\mathcal{H}$ by compressing it into $\mathcal{H}_r$:
First, we'll derive an upper on the distance between $\hat{h} \in \mathcal{H}$ and $\mathcal{H}_r$, like we've seen in class :

$$d(\hat{h}, \mathcal{H}_r) \leq \sup_{x \in X} \left\| \hat{h}(x) - h_r(x) \right\|$$
$$\leq \gamma^{N-1} \sum_{n=1}^{N} \prod_{j \in [N] \setminus \{n\}} \left\| W_j \right\|_{spectral} \| W_N - U_N V_N^T \|_{spectral}$$

Now we'll decompose W to its SVD : $W_i = \mathrm{U_i \Sigma_i V_i^T}$, where Σ_i has d or less non-zero values on its diagonal, note that the "closest" r (or less) ranked matrix would be $\mathrm{U_i \Sigma_i' V_i^T}$ where Σ_i' has the first r (or less) values of Σ_i on its diagonal, so we get:

$$\gamma^{N-1} \sum_{n=1}^{N} \prod_{j \in [N] \setminus \{n\}} \left\| W_j \right\|_{spectral} \| W_N - U_N V_N^T \|_{spectral}$$
$$= \left\| \mathrm{U_N \Sigma_N V_N^T} - \mathrm{U_N \Sigma_N' V_N^T} \right\|_{spectral} \gamma^{N-1} \sum_{n=1}^{N} \prod_{j \in [N] \setminus \{n\}} \left\| \mathrm{U_j \Sigma_j V_j^T} \right\|_{spectral}$$
$$= \left\| \mathrm{U_N (\Sigma_N - \Sigma_N') V_N^T} \right\|_{spectral} \gamma^{N-1} \sum_{n=1}^{N} \prod_{j \in [N] \setminus \{n\}} \left\| \mathrm{U_j \Sigma_j V_j^T} \right\|_{spectral}$$
$$= \sigma_N^{r+1} \gamma^{N-1} \sum_{n=1}^{N} \prod_{j \in [N] \setminus \{n\}} \sigma_j^1$$

Where σ_j^i is the i singular value of matrix j.

Now we'll put it together with the bound we have seen in class ,over all we get:

$\forall \delta \in (0,1)\ W.P. \geq 1 - \delta\ on\ S \sim D^m$:

$$L_D(\hat{h}) - L_S(\hat{h}) \leq \sqrt{\frac{(b+1)\ln(2) + \ln\left(\frac{1}{\delta}\right)}{2m}} + 2\rho d(\hat{h}, \mathcal{H}_r)$$
$$\leq \sqrt{\frac{(b+1)\ln(2) + \ln\left(\frac{1}{\delta}\right)}{2m}} + 2\rho \sigma_N^{r+1} \gamma^{N-1} \sum_{n=1}^{N} \prod_{j \in [N] \setminus \{n\}} \sigma_j^1$$

b) For simultaneously compressing for all $r \in [d]$ we'll use the bound we got above, and simply take the r that gets the minimum value, namely:

$$\min_{r \in [d]} \sqrt{\frac{(b+1)\ln(2) + \ln\left(\frac{1}{\delta}\right)}{2m}} + 2\rho \sigma_N^{r+1} \gamma^{N-1} \sum_{n=1}^{N} \prod_{j \in [N] \setminus \{n\}} \sigma_j^1$$

Radamacher complexity and norms

1. Denote $\Theta_k \coloneqq \left\{\theta | \frac{1}{2k} \leq \|\theta\|_\infty \leq \frac{1}{2}\right\}, \mathcal{H}_k \coloneqq \{h_\theta | \theta \in \Theta_k\}$
   Consider the following sequence:
   $$\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots \subseteq \mathcal{H}_\infty \coloneqq \mathcal{H}$$
   And note that $\bigcup_{k=1}^{\infty} \mathcal{H}_k = \mathcal{H}$
   then according to proposition we've seen in class:

$$\forall \delta \in (0.1)\ w.p. \geq 1-\delta\ over\ S \sim D^m : \forall k \in \mathbb{N}, h \in \mathcal{H}_k :$$

$$L_D(h) - L_S(h) \leq 2R(l \circ \mathcal{H}_k \circ S) + \sqrt{\frac{2\ln\left(\frac{2\pi^2 k^2}{3\delta}\right)}{m}}$$

Note that $R(l \circ \mathcal{H}_\Theta \circ S)$ satisfy McDiarmid's condition with $c = \frac{2}{m}$ so as we've seen, with probability $\geq 1-\delta$ it holds that

$$|R(l \circ \mathcal{H}_\Theta \circ S) - \mathbb{E}[R(l \circ \mathcal{H}_\Theta \circ S)]| \leq \sqrt{\frac{2\ln\left(\frac{4}{\delta}\right)}{m}}$$

$$\Downarrow$$

$$R(l \circ \mathcal{H}_\Theta \circ S) \leq \mathbb{E}[R(l \circ \mathcal{H}_\Theta \circ S)] + \sqrt{\frac{2\ln\left(\frac{4}{\delta}\right)}{m}}$$

Now, note that $\Theta \coloneqq \Theta_\infty$ is a hyper-cube, which means

$$1 = Volume(\Theta) = \mathbb{E}_S[R(l \circ \mathcal{H}_\Theta \circ S)]$$

Put it all together, it holds that:

$\forall \delta \in (0.1)\ w.p. \geq 1-\delta\ over\ S \sim D^m, \forall h \in \mathcal{H}:$

$$L_D(h) - L_S(h) \leq 2R(l \circ \mathcal{H} \circ S) + \sqrt{\frac{2\ln\left(\frac{2\pi^2}{3\delta}\right)}{m}}$$

$$\leq 2\left[\mathbb{E}[R(l \circ \mathcal{H} \circ S)] + \sqrt{\frac{2\ln\left(\frac{4}{\delta}\right)}{m}}\right] + \sqrt{\frac{2\ln\left(\frac{2\pi^2}{3\delta}\right)}{m}}$$

$$= 2\left[1 + \sqrt{\frac{2\ln\left(\frac{4}{\delta}\right)}{m}}\right] + \sqrt{\frac{2\ln\left(\frac{2\pi^2}{3\delta}\right)}{m}} \leq 2 + 3\sqrt{\frac{2\ln\left(\frac{2\pi^2}{3\delta}\right)}{m}}$$

PAC-Bayes

1. $Q \sim \mathcal{N}(\mu_0, \Sigma_0), P \sim \mathcal{N}(\mu_1, \Sigma_1)$ over $\mathbb{R}^r$

$$KL(Q|P) = \mathbb{E}_{x\sim Q}\left[\ln\left(\frac{Q(x)}{P(x)}\right)\right] \overset{by\ \mathrm{def}\ of\ normal\ dist}{=}$$

$$\mathbb{E}\left[\ln\left(\frac{(2\pi)^{-\frac{r}{2}}\det(\Sigma_0)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right)}{(2\pi)^{-\frac{r}{2}}\det(\Sigma_1)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right)}\right)\right] =$$

$$\mathbb{E}\left[\ln\left(\frac{\det(\Sigma_0)^{-\frac{1}{2}}}{\det(\Sigma_1)^{-\frac{1}{2}}}\right) + \ln\left(\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) + \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right)\right)\right] =$$

$$\frac{1}{2}\ln\left(\frac{\det(\Sigma_0)}{\det(\Sigma_1)}\right) + \frac{1}{2}\mathbb{E}[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)] - \frac{1}{2}\mathbb{E}[(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)] =$$

$$\frac{1}{2}\left[\ln\left(\frac{\det(\Sigma_0)}{\det(\Sigma_1)}\right) + \mathbb{E}[(x+\mu_0-\mu_0-\mu_1)^T\Sigma_1^{-1}(x+\mu_0-\mu_0-\mu_1)] - \mathbb{E}\left[trace\left((x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right)\right]\right] =$$

$$\frac{1}{2}\left[\ln\left(\frac{\det(\Sigma_0)}{\det(\Sigma_1)}\right) + \mathbb{E}[(x+\mu_0-\mu_0-\mu_1)^T\Sigma_1^{-1}(x+\mu_0-\mu_0-\mu_1)] - \mathbb{E}[trace((x-\mu_0)^T(x-\mu_0)\Sigma_0^{-1})]\right] =$$

$$\frac{1}{2}\left[\ln\left(\frac{\det(\Sigma_0)}{\det(\Sigma_1)}\right) + \mathbb{E}[(x-\mu_0)^T\Sigma_1^{-1}(x-\mu_0) + (x-\mu_0)^T\Sigma_1^{-1}(\mu_0-\mu_1) + (\mu_0-\mu_1)^T\Sigma_1^{-1}(x-\mu_0) + (\mu_0-\mu_1)^T\Sigma_1^{-1}(\mu_0-\mu_1)] - trace(\mathbb{E}[(x-\mu_0)^T(x-\mu_0)\Sigma_0^{-1}])\right] =$$

$$\frac{1}{2}\left[\ln\left(\frac{\det(\Sigma_0)}{\det(\Sigma_1)}\right) + trace(\mathbb{E}[(x-\mu_0)^T(x-\mu_0)\Sigma_1^{-1}]) + trace(\mathbb{E}[(x-\mu_0)^T\Sigma_1^{-1}(\mu_0-\mu_1)]) + trace(\mathbb{E}[(\mu_0-\mu_1)^T\Sigma_1^{-1}(x-\mu_0)]) + (\mu_0-\mu_1)^T\Sigma_1^{-1}(\mu_0-\mu_1) - trace(\mathbb{E}[(x-\mu_0)^T(x-\mu_0)]\Sigma_0^{-1})\right] =$$

(*) explanation for this equality -

$$[\mathbb{E}[(x-\mu_i)^T(x-\mu_i)] = \Sigma_i \Rightarrow \mathbb{E}[(x-\mu_i)^T(x-\mu_i)\Sigma_i^{-1}] = \mathbb{E}[\Sigma_i\Sigma_i^{-1}] = \mathbb{E}[I] = I]$$

$$= \frac{1}{2}\left[\ln\left(\frac{\det(\Sigma_0)}{\det(\Sigma_1)}\right) + trace(\Sigma_0\Sigma_1^{-1}) + trace\big(\mathbb{E}[(x-\mu_0)^T]\Sigma_1^{-1}(\mu_0-\mu_1)\big) \right.$$
$$\left. + trace((\mu_0-\mu_1)^T\Sigma_1^{-1}\mathbb{E}[(x-\mu_0)]) + (\mu_0-\mu_1)^T\Sigma_1^{-1}(\mu_0-\mu_1) - r\right]$$
$$= \frac{1}{2}\left[\ln\left(\frac{\det(\Sigma_0)}{\det(\Sigma_1)}\right) + trace(\Sigma_0\Sigma_1^{-1}) + trace\big(0\cdot\Sigma_1^{-1}(\mu_0-\mu_1)\big) \right.$$
$$\left. + trace((\mu_0-\mu_1)^T\Sigma_1^{-1}\cdot 0) + (\mu_0-\mu_1)^T\Sigma_1^{-1}(\mu_0-\mu_1) - r\right] =$$
$$= \frac{1}{2}\left[\ln\left(\frac{\det(\Sigma_0)}{\det(\Sigma_1)}\right) + trace(\Sigma_0\Sigma_1^{-1}) + (\mu_0-\mu_1)^T\Sigma_1^{-1}(\mu_0-\mu_1) - r\right]$$

■

2. In class we saw that for any dist P, we can define $Q \sim \mathcal{N}\left(\widehat{\Theta}, \hat{\sigma}^2 I\right)$ where $\widehat{\Theta} \in \mathbb{R}^r$ are the params returned by training algorithm and and $\hat{\sigma}^2$ is some variance that we fix in advance. We saw that $\forall\delta \in (0,1), w.p. \geq 1-\delta$

$$L_D(Q) \leq L_S(Q) + \sqrt{\frac{KL(Q|P) + \ln\left(\frac{2m}{\delta}\right)}{2(m-1)}}$$

We would like to define P such that the regularization that attempts to produce a solution close to at least one of a finite set of points $\{\Theta_i | i \in [k]\}$ generalizes well.

So, for each of $\{\Theta_i | i \in [k]\}$ denote $P_i \sim \mathcal{N}(\Theta_i, \sigma^2 \cdot I)$ with σ^2 being some variance that we fixed in advance.

From the lemma we proved above, it hold that

$$KL(Q|P_i) = \frac{1}{2}\left[r\cdot\frac{1}{\sigma^2}\cdot\hat{\sigma}^2 + \frac{1}{\sigma^2}\cdot\left\|\widehat{\Theta} - \Theta_i\right\|^2 - r + r\ln(\sigma^2) - r\ln(\hat{\sigma}^2)\right]$$

For $\hat{\sigma}^2 = \sigma^2$ we get

$$KL(Q|P_i) = \frac{\left\|\widehat{\Theta} - \Theta_i\right\|^2}{2\sigma^2}$$

And so, it holds that $w.p. \geq 1-\delta$:

$$L_D(Q) \leq L_S(Q) + \sqrt{\frac{\frac{\left\|\widehat{\Theta} - \Theta_i\right\|^2}{2\sigma^2} + \ln\left(\frac{2m}{\delta}\right)}{2(m-1)}}$$

Choosing $\delta^* = \frac{\delta}{k}$ we get:

$$\mathbb{P}\left(\exists i, L_D(Q) \geq L_S(Q) + \sqrt{\frac{\frac{\left\|\widehat{\Theta} - \Theta_i\right\|^2}{2\sigma^2} + \ln\left(\frac{2m}{\delta^*}\right)}{2(m-1)}} \right)$$

$$\leq \sum_{i=1}^{k} \mathbb{P}\left(L_D(Q) \geq L_S(Q) + \sqrt{\frac{\frac{\left\|\widehat{\Theta} - \Theta_i\right\|^2}{2\sigma^2} + \ln\left(\frac{2m}{\delta^*}\right)}{2(m-1)}} \right) \leq k\delta^* = \delta$$

So w.p. $\geq 1 - \delta$ it holds:

$$\forall i, L_D(Q) \geq L_S(Q) + \sqrt{\frac{\frac{\left\|\widehat{\Theta} - \Theta_i\right\|^2}{2\sigma^2} + \ln\left(\frac{2m}{\delta^*}\right)}{2(m-1)}}$$

■

Part 3: Implicit Pegularization

1. Proposition: With the notations and setting established in class, suppose we minimize $L_S(w)$ by initializing $w^{(0)} = 0$, and producing iterates $w^{(1)}, w^{(2)}, w^{(3)}, \ldots$ via iterative algorithm in which every update $w^{(t+1)} - w^{(t)}$ lies in $span\{\nabla l_{(x_i,y_i)}(w) : i \in [m], w \in R^d\}$ (this includes as special cases gradient descent, stochastic gradient descent and momentum). Assume convergence to global optimum, i.e. to a solution with zero loss. The, this global optimum is the one with minimum (Euclidean) norm.

   Generalize this result by proving that if the zero initialization is replaced by an arbitrary initialization $w^{(0)} = a \in R^d$, then the sub-optimality of the obtained norm (i.e. the extent to which it is larger than min norm across all global optima) is $\leq ||P_\perp a||$, where $P_\perp : R^d \to R^d$ stands for projection onto the orthogonal complement of $span\{x_i\}_{i=1}^m$.

   As we saw in class, for any $i \in [m]$ and $w \in R^d$: $\nabla l_{(x_i,y_i)}(w) = (< x_i, w > - y_i) \cdot x_i$. This implies that for every $t \in N \cup \{0\}$, $w^{(t+1)} - w^{(t)} \in span\{x_i\}_{i=1}^m$.

   Now since we assume $w^{(0)} = a \in R^d$, we get that $\forall t \in N: w^{(t)} \in a + span\{x_i\}_{i=1}^m$.

   $a + span\{x_i\}_{i=1}^m$ is topologically closed in R^d, therefore $w^{(\infty)} \coloneqq \lim_{t\to\infty} w^{(t)}$ must lie in $a + span\{x_i\}_{i=1}^m$, i.e. $\exists r. w^{(\infty)} = a + x_r$.

   Since $L_S\left(w^{(\infty)}\right) = 0$:

$$X^T w^{(\infty)} = y \to X^T(a + x_r) = y \to r = (X^T X)^{-1} y - (X^T X)^{-1} X^T a$$
$$\to w^{(\infty)} = a + x_r = a + X(X^T X)^{-1} y - X(X^T X)^{-1} a$$

   denote $a = a_{||} + a_\perp$ where:

   - $a_{||}$ is the projection of a onto $span\{x_i\}_{i=1}^m$
   - $a_\perp$ is the projection of a onto $\perp span\{x_i\}_{i=1}^m$

   So $w^{(\infty)} = \underbrace{a_\perp}_{\coloneqq w_\perp} + \underbrace{a_{||} + X(X^T X)^{-1} y - X(X^T X)^{-1} a}_{\coloneqq w_{||}} \to y = X^T w^{(\infty)} = \underbrace{X^T w_\perp}_{=0} + X^T w_{||}$

   Meaning $X^T w_{||}$ is also a global optimum (as is $X^T w^{(\infty)}$), which we showed in class has the minimum norm (denote $\left||w|\right|^*$), therefore:

$$\left||w^{(\infty)}|\right| = \left||w_\perp + w_{||}|\right| \leq ||w_\perp|| + \left||w_{||}|\right| = ||a_\perp|| + \left||w|\right|^* = \left||P_\perp a|\right| + \left||w|\right|^* \quad \blacksquare$$