# Foundations of Deep Learning – Homework Assignment #4

Adi Album & Tomer Epshtein

**Part 1:**

Experiment:

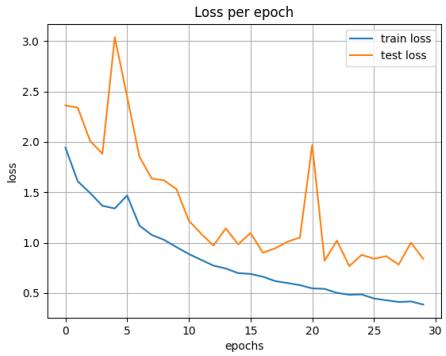
Demonstrate the four empirical postulates we relied on in class when rationalizing about generalization in deep learning.

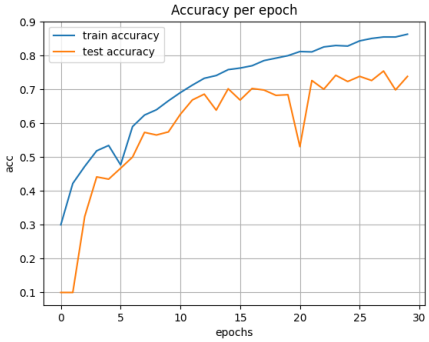
Results:

We chose to work with the CIFAR10 dataset with the common MobileNetV2 architecture.

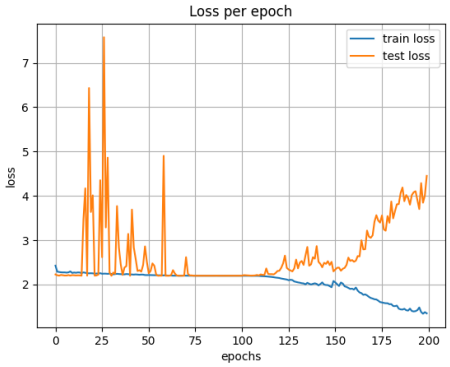
Here are our results:

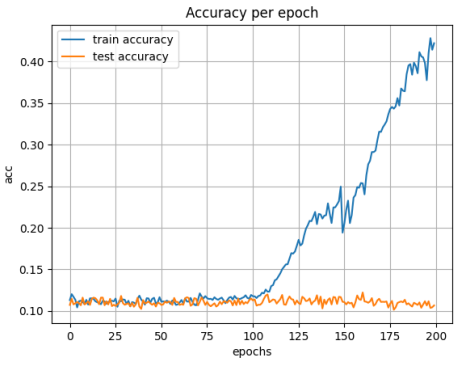
1. Standard NNs (=MobileNetV2) trained by a standard optimization algorithm (=Adam) on a standard dataset (=CIFAR10) generalize well without any explicit regularization. Even though # of learned parameters >> # of training examples.



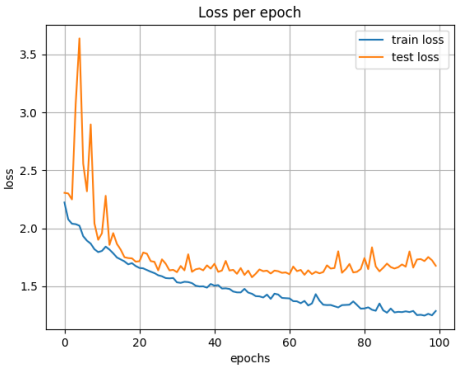



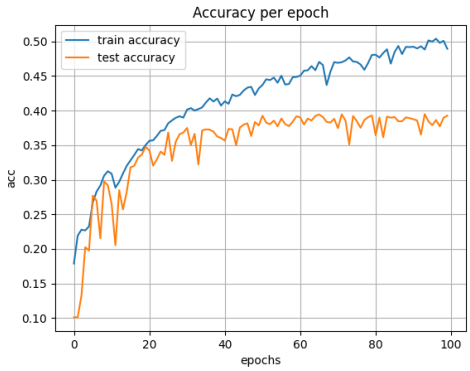

2. In the regime of (1.), the training error of learned hypothesis $\approx 0$. This remains the case even if we replace the training examples by any training set of the same size. We demonstrated this by replacing CIFAR10 dataset with random images (with every pixel randomly samples from a uniform 0,1 distribution) and every image is uniformly at random labeled 0-9.



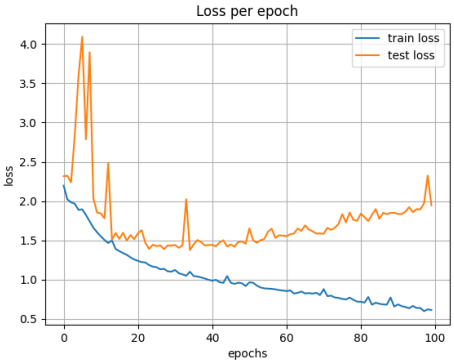



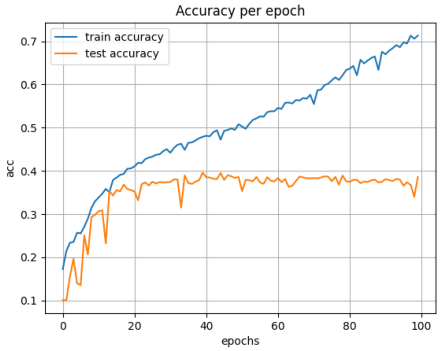

3. In the regime of (1.), with half the training set being CIFAR10 and the second half is replaced by random data, the test error of the learned hypothesis (whose training error is of course $\approx 0$) is far better than trivial.




4. In the regime of (1.), with half of the training being CIFAR10 and the second half replaced by adversarially labeled CIFAR10, test error significantly deteriorates



## Compression

$$H = \{ \mathbb{R}^d \ni x \mapsto y = W_N \sigma(\dots W_2 \sigma(W_1 x) \dots) \in \mathbb{R}^d : W_n \in \mathbb{R}^{d,d}, n \in [N] \}$$

Assume $\sigma$ is $\gamma$-Lipschitz and $\sigma(0) = 0$, $X = \{x \in \mathbb{R}^d, \|x\| \le 1\}$

For $r \in [d]$ Let $H_r$ to be the hypotheses space corresponding to the same network above when its weight matrices are constrained to have rank $r$ or less, i.e:

$$H_r = \{ x \to y = W_N V_N^T \sigma(U_{N-1} V_{N-1}^T \cdots \sigma(U_1 V_1^T x) \dots) : U_n, V_n \in \mathbb{R}^{d,r} \; n \in [N] \}$$

Assume that each of the $2Ndr$ (parameters) representing $H_r$ is stored in memory using $b$ bits.

Given a loss $\ell : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$ s.t $\forall y, \hat{y}, y' \; |\ell(y, \hat{y}) - \ell(y, y')| \le \rho \cdot \|\hat{y} - y'\|$

we would like to derive generalization bounds for $H$.

(a) Fix $r \in [d]$ and derive a generalization bound for $H$ by compressing it into $H_r$.

(b) Derive a generalization bound for $H$ by simultaneously compressing it into $H_r$ for all $r \in [d]$

## Solution (a)

Let $r \in [d]$.

By assumption each of the $2Ndr$ parameters representing $H_r$ using $b$ bits, thus $|H_r| \le \underbrace{2^b}_{\#\text{options for the first param}} \cdot \dots \cdot \underbrace{2^b}_{\#\text{options for the } 2Ndr \text{ param}} = 2^{b \cdot 2Ndr}$

In addition, $\ell(\cdot,\cdot)$ is $\gamma$-Lipschitz, therefore we can use the theorem we proved in class in order to get:

For $\hat{h}$ the returned hypothesis from the algorithm $\forall \delta \in (0,1)$ w.p $\ge 1 - \delta$ over $S \sim D^m$:

$$L_D(\hat{h}) - L_S(\hat{h}) \le \sqrt{\frac{(b \cdot 2Ndr + 1)\ln(2) + \ln(\frac{1}{\delta})}{2m}} + 2\rho \cdot d(\hat{h}, H_r)$$

We would like to bound the term $d(\hat{h}, H_r)$.

Note that we proved in class the following (under the same assumptions except the rank of the approximation matrices in class it was rank=1 and now rank=r):

$$d(\hat{h}, h_r) \leq \gamma^{N-1} \cdot \sum_{n=1}^{N} \prod_{j \in [N] \setminus n} \|W_j\|_{spectral} \cdot \|W_n - W_n'\|_{spectral}$$

For $\hat{h}$ with weighted matrices $W_N, \ldots, W_1$ and $h_r$ with weighted matrices $W_N', \ldots, W_1'$ with rank $\leq r$ the closest approximations to $W_N, \ldots, W_1$.

In the proof in class the only place we used that the approximations were to 1-rank matrix was when explaining $\|W_n'\|_{spectral} = \|W_n\|_{spectral} \; \forall n$

Note, that this equation holds for approximations to rank $r \geq 1$ too, since if we denote $W_n$ as spectral decomposition $W_n = \sum_{i=1}^{d} \sigma_i u_i v_i^T$ where $|\sigma_1| \geq |\sigma_2| \geq \ldots \geq |\sigma_d|$ then the best rank-$r$ approximation to $W_n$ is $W_n' = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, thus we can see clearly that $\|W_n'\|_{spectral} = |\sigma_1| = \|W_n\|_{spectral}$ and thus the bound denote above holds for each $n \in [d]$.

In addition $\|W_n - W_n'\|_{spectral} = \| \sum_{i=1}^{d} \sigma_i u_i v_i^T - \sum_{i=1}^{r} \sigma_i u_i v_i^T \|_{sp} = \| \sum_{i=r+1}^{d} \sigma_i u_i v_i^T \|_{sp} = |\sigma_{r+1}|$

$$= |\sigma_{r+1}(W_n)|$$

$\Rightarrow$ In total we'll get $\forall r \in (0,1)$ w.p $\geq 1-\delta$ over $S \sim D^m$:

$$L_D(\hat{h}) - L_S(\hat{h}) \leq \sqrt{\frac{(b^2 N(r+1)) \ln(2) + \ln(\frac{1}{\delta})}{2m}} + 2 \cdot \delta \cdot \left[ \gamma^{N-1} \sum_{n=1}^{N} \prod_{j \in [N] \setminus n} \|W_j\|_{spec} \cdot |\sigma_{r+1}(W_n)| \right]$$

□

Solution (b):

We'll use the bound from (a) with union bound in order to get bound for H.

From (a) we've got that $\forall \delta \in (0,1)$ w.p $\geq 1-\frac{\delta}{d}$ over $S \sim D^m$ and fixed $1 \leq r \leq d$:

$$L_D(\hat{h}) - L_S(\hat{h}) \leq \sqrt{\frac{(b \cdot 2Ndr+1)\ln(2) + \ln\left(\frac{1}{\frac{\delta}{d}}\right)}{2(m-1)}} + 2s \cdot \left[\gamma^{N-1} \sum_{n=1}^{N} \prod_{j \neq n} \|W_j\|_{sp} \cdot |\sigma_{r+1}(W_n)|\right]$$

$$= \sqrt{\frac{(b \cdot 2Ndr+1)\ln(2) + \ln\left(\frac{d}{\delta}\right)}{2(m-1)}} + 2s \cdot \left[\gamma^{N-1} \sum_{n=1}^{N} \prod_{j \neq n} \|W_j\|_{sp} \cdot |\sigma_{r+1}(W_n)|\right]$$

We will sign $A_r = \sqrt{\frac{(b \cdot 2Ndr+1)\ln(2) + \ln\left(\frac{d}{\delta}\right)}{2(m-1)}}$, $B_r = 2 \cdot s \cdot \left[\gamma^{N-1} \sum_{n=1}^{N} \prod_{j \neq n} \|W_j\|_{sp} \cdot |\sigma_{r+1}(W_n)|\right]$

Thus $P_r\left(L_D(\hat{h}) - L_S(\hat{h}) \geq A_r + B_r\right) \leq \frac{\delta}{d} \quad \forall 1 \leq r \leq d$

$$\Rightarrow P_r\left(\exists r\, L_D(\hat{h}) - L_S(\hat{h}) \geq A_r + B_r\right) \underset{\text{union bound}}{\leq} \sum_{r=1}^{d} P_r\left(L_D(\hat{h}) - L_S(\hat{h}) \geq A_r + B_r\right) \leq \sum_{r=1}^{d} \frac{\delta}{d} = \delta$$

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

$$L_D(\hat{h}) - L_S(\hat{h}) \leq \min_{1 \leq r \leq d} (A_r + B_r) = \min_{1 \leq r \leq d} \left\{ \sqrt{\frac{(b \cdot 2Ndr+1)\ln(2) + \ln\left(\frac{d}{\delta}\right)}{2(m-1)}} + 2s \left[\gamma^{N-1} \sum_{n=1}^{N} \prod_{j \neq n} \|W_j\|_{sp} \cdot |\sigma_{r+1}(W_n)|\right] \right\}$$

Note that as r gets bigger: $A_r$ gets bigger and $B_r$ gets lower (better approximation to the matrices but another case of more weighted parameters)

□

# Foundations of Deep Learning – Homework Assignment #4

Adi Album & Tomer Epshtein

## Part 2: (2)

### Question:

Let $H = \{h_\theta : \mathcal{X} \to \mathcal{Y} : \theta \in \mathbb{R}^p, \|\theta\|_\infty \leq 0.5\}$ be a hypothesis space corresponding to a neural network with $p$ parameters bounded in $[-0.5, 0.5]$. For any subset $\Theta \subseteq \{\theta \in \mathbb{R}^p : \|\theta\|_\infty \leq 0.5\}$, denote $\mathcal{H}_\Theta := \{h_\theta : \theta \in \Theta\}$. Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$ and a training set $S = \{(x_i, y_i)\}_{i=1}^m \subseteq \mathcal{X} \times \mathcal{Y}$, the Radamacher complexity of $\mathcal{H}_\Theta$ is defined to be:

$$R(\ell \circ \mathcal{H}_\Theta \circ S) := \frac{1}{m}\mathbb{E}_\xi\left[\sup_{v \in \ell \circ \mathcal{H}_\Theta \circ S} \sum_{i=1}^m \xi_i v_i\right]$$

Where:

- $\xi$ is short for $\xi_1, \dots, \xi_m \overset{i.i.d}{\sim} \begin{cases} +1, & w.p\ 0.5 \\ -1, & w.p\ 0.5 \end{cases}$
- $\ell \circ \mathcal{H}_\Theta \circ S = \left\{\left(\ell(y_1, h(x_1)), \ell(y_2, h(x_2)), \dots, \ell(y_m, h(x_m))\right) : h \in \mathcal{H}_\Theta\right\} \subseteq \mathbb{R}^m$

Assume that

$$\mathbb{E}_S[R(\ell \circ \mathcal{H}_\Theta \circ S)] = Volume(\Theta) := \int_{\theta \in \mathbb{R}^p} 1_{[\theta \in \Theta]} d\theta$$

Assume also that the implicit regularization of optimization leads to solutions with high $\|\cdot\|_\infty$, i.e., to:

$$h_{\hat{\theta}} \in \mathcal{H}, \hat{\theta} \in argmax_{\theta \in \mathbb{R}^p, \|\theta\|_\infty \leq 0.5} \|\theta\|_\infty \; s.t \; \theta \; minimizes \; training \; loss$$

Derive a generalization bound for $\mathcal{H}$ that takes advantage of our knowledge on the implicit regularization, i.e. under which learned solutions with high $\|\cdot\|_\infty$ ensure small generalization gap.

### Proof:

Define $\Theta^{(c)} := \left\{\theta \in \mathbb{R}^p : \frac{1}{2} - c <= \|\theta\|_\infty \leq \frac{1}{2}\right\}$ for $c \in \left[0, \frac{1}{2}\right]$.

Let $\epsilon > 0$ be a small positive real value such that for some $t \in \mathbb{N}$: $t\epsilon = \frac{1}{2}$. We define a series $\Theta_1, \Theta_2, \Theta_3, \dots$ by:

- For all $i \leq t$: $\Theta_i := \Theta^{(i \cdot \epsilon)}$
- For all $i > t$: $\Theta_i := \Theta$

We have $\Theta_1 \subseteq \Theta_2 \subseteq \cdots$. For each $i \in \mathbb{N}$, $\Theta_i$ defines $\mathcal{H}_{\Theta_i}$ and this produces a series of subsets: $\mathcal{H}_{\Theta_1} \subseteq \mathcal{H}_{\Theta_2} \subseteq \cdots$

1. For every $k \in \mathbb{N}$ such that $k\epsilon \leq \frac{1}{2}$:

   $$\mathbb{E}_S\left[R\left(\ell \circ \mathcal{H}_{\Theta_k} \circ S\right)\right] = Volume(\Theta_k) = \left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p$$

   $R\left(\ell \circ \mathcal{H}_{\Theta_k} \circ S\right)$ is a non-negative random variable, so by Markov's Inequality:
   $\forall a > 0$:

   $$\Pr_S\left(R\left(\ell \circ \mathcal{H}_{\Theta_k} \circ S\right) \geq a\right) \leq \frac{\mathbb{E}_S\left[R\left(\ell \circ \mathcal{H}_{\Theta_k} \circ S\right)\right]}{a} = \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{a} \overset{?}{\leq} \frac{\delta}{2}$$

   We want to choose $a$ such that $(?)$ holds
   I.e.

   $$\frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{a} \overset{?}{\leq} \frac{\delta}{2}$$

   Choose:

   $$a = \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/2}$$

   So we have

   $$\Pr_S\left(R\left(\ell \circ \mathcal{H}_{\Theta_k} \circ S\right) \geq \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/2}\right) \leq \frac{\delta}{2}$$

   Or by viewing complement:

   $$\Pr_S\left(R\left(\ell \circ \mathcal{H}_{\Theta_k} \circ S\right) < \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/2}\right) > 1 - \frac{\delta}{2}$$

2. By proposition proved in class,
   $\forall k \in \mathbb{N}, \delta \in (0,1)$ w.p. $\geq 1 - \frac{\delta}{2}$ over $S \sim D^m$:

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

   Let $\delta \in (0,1), k \in \mathbb{N}$.

$$\Pr\left( \forall h \in \mathcal{H}_k: L_D(h) - L_S(h) \leq 2R(\ell \circ \mathcal{H}_k \circ S) + \sqrt{\frac{2 \cdot ln\left(\frac{2\pi^2}{3} k^2 \frac{2}{\delta}\right)}{m}} \right) \geq 1 - \frac{\delta}{2}$$

3. Reminder: For any two probabilistic events $A$ and $B$ we have

$$\Pr(A \cap B) \geq \Pr(A) + \Pr(B) - 1$$

Let $k \in \mathbb{N}$ such that $k\epsilon \leq \frac{1}{2}$. Let's look at:

$$\Pr\left(\left( \forall h \in \mathcal{H}_k: L_D(h) - L_S(h) \leq 2R(\ell \circ \mathcal{H}_k \circ S) + \sqrt{\frac{2 \cdot ln\left(\frac{2\pi^2}{3} k^2 \frac{2}{\delta}\right)}{m}} \right) \cap \left( R\left(\ell \circ \mathcal{H}_{\Theta_k} \circ S\right) < \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/2} \right)\right)$$

- On the one hand:

$$\Pr\left(\left( \forall h \in \mathcal{H}_k: L_D(h) - L_S(h) \leq 2R(\ell \circ \mathcal{H}_k \circ S) + \sqrt{\frac{2 \cdot ln\left(\frac{2\pi^2}{3} k^2 \frac{2}{\delta}\right)}{m}} \right) \cap \left( R\left(\ell \circ \mathcal{H}_{\Theta_k} \circ S\right) < \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/2} \right)\right) \overset{(3)}{\geq}$$

$$\geq \Pr\left( \forall h \in \mathcal{H}_k: L_D(h) - L_S(h) \leq 2R(\ell \circ \mathcal{H}_k \circ S) + \sqrt{\frac{2 \cdot ln\left(\frac{2\pi^2}{3} k^2 \frac{2}{\delta}\right)}{m}} \right) + \Pr\left( R\left(\ell \circ \mathcal{H}_{\Theta_k} \circ S\right) < \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/2} \right) - 1 \overset{(1)\&(2)}{\geq}$$

$$\geq \left(1 - \frac{\delta}{2}\right) + \left(1 - \frac{\delta}{2}\right) - 1 = 1 - \delta$$

- On the other hand:

$$\Pr\left(\left( \forall h \in \mathcal{H}_k: L_D(h) - L_S(h) \leq 2R(\ell \circ \mathcal{H}_k \circ S) + \sqrt{\frac{2 \cdot ln\left(\frac{2\pi^2}{3} k^2 \frac{2}{\delta}\right)}{m}} \right) \cap \left( R\left(\ell \circ \mathcal{H}_{\Theta_k} \circ S\right) < \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/2} \right)\right) \leq$$

$$\leq \Pr\left( \forall h \in \mathcal{H}_k: L_D(h) - L_S(h) < \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/4} + \sqrt{\frac{2 \cdot ln\left(\frac{2\pi^2}{3} k^2 \frac{2}{\delta}\right)}{m}} \right)$$

Bringing it all together, we have:

$\forall \delta \in (0,1), \forall \epsilon > 0 \; and \; k \in \mathbb{N}$ such that $k\epsilon \leq \frac{1}{2}$, w.p. $\geq 1 - \delta$

$$h \in \mathcal{H}_{\Theta_k}: L_D(h) - L_S(h) < \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/4} + \sqrt{\frac{2 \cdot ln\left(\frac{2\pi^2}{3} k^2 \frac{2}{\delta}\right)}{m}}$$

Therefor, solutions with high $\| \cdot \|_\infty$ are hypothesis' $h \in \mathcal{H}_k$ with small $k$, yielding small generalization gaps.

# PAC - Bayes

## Question 1

Consider two multivariate Gaussian distributions over $\mathbb{R}^r$ - $N(\mu_0, \Sigma_0)$ and $N(\mu_1, \Sigma_1)$ where $\Sigma_0$ and $\Sigma_1$ are non-singular (positive definite). Then it holds that:

$$KL(N(\mu_0,\Sigma_0) \,||\, N(\mu_1,\Sigma_1)) = \frac{1}{2}\left[ tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1-\mu_0)^T \Sigma_1^{-1}(\mu_1-\mu_0) - r + \ln\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)\right]$$

## Solution

First, let's sign the two distributions as:
$P = N(\mu_0, \Sigma_0)$ and $Q = N(\mu_1, \Sigma_1)$.

By definition: $KL(P||Q) = \mathbb{E}_P\left[\ln\left(\frac{P}{Q}\right)\right]$.

Remember that the density function of a multivariate Gaussian distribution $N(\mu, \Sigma)$ is:

$$p(x) = \frac{1}{(2\pi)^{\frac{r}{2}}(\det\Sigma)^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

$\Rightarrow$

$$\frac{P(x)}{Q(x)} = p(x)\cdot\frac{1}{Q(x)} = \frac{1}{(2\pi)^{\frac{r}{2}}(\det\Sigma_0)^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)} \cdot \frac{(2\pi)^{\frac{r}{2}}(\det\Sigma_1)^{\frac{1}{2}}}{e^{-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)}}$$

$$= \frac{(\det\Sigma_1)^{\frac{1}{2}}}{(\det\Sigma_0)^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) - \left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right)}$$

$$= \left(\frac{\det\Sigma_1}{\det\Sigma_0}\right)^{\frac{1}{2}} \cdot e^{\frac{1}{2}\cdot\left[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right]}$$

Thus

$$KL(P\|Q) = \mathbb{E}_P\left[\ln\left(\frac{f}{g}\right)\right]$$

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

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

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

linearity of $\mathbb{E}(\cdot)$ ↘

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

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

claim

For each $x\in\mathbb{R}^k$, $A\in\mathbb{R}^{k\times k}$: $\quad x^TAx = tr(Axx^T)$

proof

$$x^TAx = (x_1\cdots x_k)\begin{pmatrix}|\\Ax\\|\end{pmatrix} = (x_1\cdots x_k)\begin{pmatrix}(Ax)_1\\\vdots\\(Ax)_k\end{pmatrix} = (x_1\cdots x_k)\begin{pmatrix}\sum_{j=1}^k a_{1j}x_j\\\vdots\\\sum_{j=1}^k a_{kj}x_j\end{pmatrix} =$$

$$= \sum_{j=1}^k a_{1j}x_1x_j + \cdots + \sum_{j=1}^k a_{kj}x_kx_j =$$

$$= \sum_{j=1}^k a_{1j}(xx^T)_{j1} + \cdots + \sum_{j=1}^k a_{kj}(xx^T)_{jk} =$$

$$= (Axx^T)_{11} + \cdots + (Axx^T)_{kk} = tr(Axx^T)$$

□

Continue next page...

Thus:

$$KL(P||Q) = \frac{1}{2}\left[\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) + E_p[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)] - E_p[(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)]\right]$$

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

$tr(E[A]) = E[tr(A)]$ by linearity of $E[\cdot]$ →

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

$$= \frac{1}{2}\Big[\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) + tr\Big(\Sigma_1^{-1}E_p[xx^T - \mu_1x^T - x\mu_1^T + \mu_1\mu_1^T]\Big)$$
$$- tr\left(\Sigma_0^{-1}E_p[(x-\mu_0)(x-\mu_0)^T]\right)\Big] \quad (=)$$

By definition, $P = N(\mu_0, \Sigma_0)$ thus

$E_p[x] = \mu_0$ and $E_p[(x-\mu_0)(x-\mu_0)^T] = \Sigma_0$

and $E_p[xx^T] = \Sigma_0 + \mu_0\mu_0^T$, so:

$$(=)\ \frac{1}{2}\Big[\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) + tr\Big(\Sigma_1^{-1}\big(\Sigma_0 + \mu_0\mu_0^T - \mu_1\mu_0^T - \mu_0\mu_1^T + \mu_1\mu_1^T\big)\Big)$$
$$\rightarrow tr\left(\Sigma_0^{-1}\Sigma_0\right)\Big]$$

$$= \frac{1}{2}\Big[\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) + tr\left(\Sigma_1^{-1}\Sigma_0\right) + tr\Big(\Sigma_1^{-1}\big(\mu_0\mu_0^T - \mu_1\mu_0^T - \mu_0\mu_1^T + \mu_1\mu_1^T\big)\Big)$$
$$\rightarrow tr\left(I_r\right)\Big]$$

$x^TAx = Axx^T$ by claim we proved ←

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

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

$\Rightarrow$

$$KL(N(\mu_0,\Sigma_0)||N(\mu_1,\Sigma_1)) = \frac{1}{2}\left[\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) - r + tr\left(\Sigma_1^{-1}\Sigma_0\right) + (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0)\right]$$

as required.

□

## Question 2

Suppose we know that the implicit regularization of optimization tends to flat minima, but not of low norm. Instead, it attempts to produce a solution close to at lease one of a finite set of points $\{\theta_1, \ldots, \theta_k\}$.
Using PAC-BAYES derive a generalization bound that accounts for the latter implicit regularization.

## Solution

For $1 \le i \le k$ we define $P_i$ as the Gaussian distribution $N(\theta_i, \sigma^2 I)$ ($\sigma^2 I$ is the same variance to all $\{P_i\}_{i=1}^k$)

Now, the distribution $Q$ will be $N(\hat{\theta}, \bar{\sigma}^2 I)$ where $\hat{\theta} \in \mathbb{R}^r$ are the params returned by the training algorithm and $\bar{\sigma}^2$ is some variance we fix in advance.

By the lemma we proved in question 1:

$$\forall 1 \le i \le k \quad KL(Q \| P_i) = \frac{1}{2}\left( r \cdot \frac{1}{\sigma^2} \cdot \bar{\sigma}^2 + \frac{1}{\sigma^2} \|\hat{\theta} - \theta_i\|^2 - r + r \ln(\sigma^2) - r \ln(\bar{\sigma}^2) \right)$$

As we explained in class, fixing $\hat{\theta}$ and minimizing over $\bar{\sigma}^2$ will yield to $\bar{\sigma}^2 = \sigma^2$ and $Q = N(\hat{\theta}, \sigma^2 I)$ and:

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

Let $\delta \in (0,1)$. By the theorem from class, w.p $\ge 1 - \frac{\delta}{k}$ over $S \sim D^m$:

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

Thus, we've got for every $1 \le i \le K$:

w.p $\ge 1 - \frac{\delta}{K}$ over $S \sim D^m$:

$$L_D(Q) - L_S(Q) \le \sqrt{\frac{\frac{1}{2\sigma^2}\|\hat{\theta} - \theta_i\|^2 + \ln\left(\frac{2mK}{\delta}\right)}{2(m-1)}}$$

$$\Rightarrow \quad P_r\left( L_D(Q) - L_S(Q) \ge \sqrt{\frac{\frac{1}{2\sigma^2}\|\hat{\theta} - \theta_i\|^2 + \ln\left(\frac{2mK}{\delta}\right)}{2(m-1)}} \right) \le \frac{\delta}{K}$$

$$\Rightarrow \quad P_r\left( \exists i \; L_D(Q) - L_S(Q) \ge \sqrt{\frac{\frac{1}{2\sigma^2}\|\hat{\theta} - \theta_i\|^2 + \ln\left(\frac{2mK}{\delta}\right)}{2(m-1)}} \right) \quad (*)$$

union bound $\rightarrow$ $(*) \le \sum_{i=1}^{K} P_r\left( L_D(Q) - L_S(Q) \ge \sqrt{\frac{\frac{1}{2\sigma^2}\|\hat{\theta} - \theta_i\|^2 + \ln\left(\frac{2mK}{\delta}\right)}{2(m-1)}} \right)$

$$\le \sum_{i=1}^{K} \frac{\delta}{K} = \delta$$

$\Rightarrow$ w.p $\ge 1 - \delta$ over $S \sim D^m$:

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

$$= \sqrt{\frac{\frac{1}{2\sigma^2} \cdot \min_{1 \le i \le K}\|\hat{\theta} - \theta_i\|^2 + \ln\left(\frac{2mK}{\delta}\right)}{2(m-1)}}$$

Since $\hat{\theta}$ tends to ~~[illegible]~~ be close to at least one of $\{\theta_i\}_{i=1}^{K}$ we'll get that $\min_{1 \le i \le K}\|\hat{\theta} - \theta_i\|^2$ will be with low value.

In addition by assumption the optimization tends to flat minima which means $L_S(Q)$ tends to be low. Thus, in total the generalization bound on $L_D(Q)$ tends to be low as required.

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

□

Part 3 Implicit Regularization / linear regression

Question 1

Prove the following proposition.

Proposition:

With the notations and setting established in class suppose we minimize $L_S(w)$ by initializing $w^{(0)} = a \in \mathbb{R}^d$ and producing iterates $w^{(1)}, w^{(2)}, \ldots$ via iterative algorithm in which every update $w^{(t+1)} - w^{(t)} \in \text{span}\{\nabla \ell_{(x_i,y_i)}(w) : i \in [m], w \in \mathbb{R}^d\}$.

Assume convergence to global min with zero loss.

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 $P_\perp : \mathbb{R}^d \to \mathbb{R}^d$ stands for projection onto the orthogonal complement of span $\{x_i\}_{i=1}^m$.

Proof

Denote by $W^* = \underset{\substack{w \in \mathbb{R}^d \text{ global min} \\ L_S(w) = 0}}{\arg\min} \|w\|$ the solution with the min norm across all global optima.

Lemma 1

The solution under the settings in the proposition is

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

With the markings above and as far as lemma 1 is true we have to prove that

$$\|\tilde{w}\| \leq \|w^*\| + \|P_\perp a\|$$

in order to finish the question.

## Proof of lemma 1

First, we'll prove that for every $t \in \mathbb{N}$ $w^t \in \underbrace{a + \text{span}\{x_i\}_{i=1}^m}_{\text{affine subspace}}$ by induction.

**base:** $w^0 = a = a + \sum_{i=1}^m 0 \cdot x_i \in a + \text{span}\{x_i\}_{i=1}^m$

**Step:** Suppose $w^t \in a + \text{span}\{x_i\}_{i=1}^m$. We'll prove that $w^{t+1} \in a + \text{span}\{x_i\}_{i=1}^m$

$w^t \in a + \text{span}\{x_i\}_{i=1}^m \Rightarrow w^t = a + \sum_{i=1}^m \alpha_i x_i$ for $\{\alpha_i\}_{i=1}^m$ scalars.

Note that for any $i \in [m]$ and $w \in \mathbb{R}^d$

$$\nabla \ell(w, x_i, y_i) = (x_i^T w - y_i) \cdot x_i$$

This implies that for every $t \in \mathbb{N}$

$$w^{t+1} - w^t \in \text{span}\{x_i\}_{i=1}^m$$

(Since we know that $w^{t+1} - w^t \in \text{span}\{\nabla \ell_{(x_i,y_i)}(w) \mid i \in [m], w \in \mathbb{R}^d\}$)

$$\Rightarrow \quad w^{t+1} - w^t = \sum_{i=1}^m \beta_i x_i \quad \text{for some } \{\beta_i\}_{i=1}^m \text{ scalars.}$$

$$\Rightarrow \quad w^{t+1} - \left(a + \sum_{i=1}^m \alpha_i x_i\right) = \sum_{i=1}^m \beta_i x_i$$

$$\Rightarrow \quad w^{t+1} = a + \sum_{i=1}^m (\alpha_i + \beta_i) x_i \quad \Rightarrow w^{t+1} \in a + \text{span}\{x_i\}_{i=1}^m$$

□

Hence we get for every $t \in \mathbb{N}$ $w^t \in a + \text{span}\{x_i\}_{i=1}^m$.

Since the subspace $a + \text{span}\{x_i\}_{i=1}^m$ is topologically closed

$$\tilde{w} = \lim_{t \to \infty} w^t \in a + \text{span}\{x_i\}_{i=1}^m$$

$$\Rightarrow \quad \tilde{w} = a + \sum_{i=1}^m r_i x_i \quad \text{for some scalars } \{r_i\}_{i=1}^m$$

$$\Rightarrow \quad \tilde{w} = a + Xr \quad \text{for } r = \begin{pmatrix} r_1 \\ \vdots \\ r_m \end{pmatrix}$$

(continue of the proof next page)

We've got $\tilde{w} = a + Xr$.

Since the convergence is to global min with zero loss we get

$X^T\tilde{w} = y \Rightarrow X^T(a + Xr) = y$

$\Rightarrow X^Ta + X^TXr = y$

$\Rightarrow X^TXr = y - X^Ta$

$\text{rank}(X^TX) = m \Rightarrow r = (X^TX)^{-1}(y - X^Ta)$

$\Rightarrow \tilde{w} = a + X(X^TX)^{-1}(y - X^Ta)$

as required. □

Now, remember that at class we proved that the solution $w^* = X(X^TX)^{-1}y$ is the one with minimal norm.

Thus we have to show that:

$$\|a + X(X^TX)^{-1}(y - X^Ta)\| = \|\tilde{w}\| \leq \|w^*\| + \|P_\perp a\| = \|X(X^TX)^{-1}y\| + \|P_\perp a\|$$

We'll show that:

$$\|a + X(X^TX)^{-1}(y - X^Ta)\| = \|a + X(X^TX)^{-1}y - X(X^TX)^{-1}X^Ta\|$$

$$\leq \|X(X^TX)^{-1}y\| + \|a - X(X^TX)^{-1}X^Ta\|$$

△ inequality

$$= \|X(X^TX)^{-1}y\| + \|P_\perp a + P_{||}a - X(X^TX)^{-1}X^T(P_\perp a + P_{||}a)\|$$

$a = P_{||}a + P_\perp a$

where $P_{||}a \in \text{span}\{x_i\}$ and $P_\perp a$ is the projection as described in the question.

Notice that by definition $\langle x_i, P_\perp a\rangle = 0 \ \forall i \Rightarrow X^TP_\perp a = \begin{pmatrix}0\\ \vdots \\ 0\end{pmatrix}$

thus:

$$\Leftarrow \|X(X^TX)^{-1}y\| + \|P_\perp a + P_{||}a - X(X^TX)^{-1}X^TP_{||}a\| \Leftarrow$$

$P_{||}a \in \text{span}\{x_i\}_{i=1}^m$, thus $P_{||}a = X \cdot b$ for some $b = \begin{pmatrix}b_1\\ \vdots \\ b_m\end{pmatrix}$

$$\Leftarrow \|X(X^TX)^{-1}y\| + \|P_\perp a + Xb - X\underline{(X^TX)^{-1}X^TX}b\|$$

$$= \|X(X^TX)^{-1}y\| + \|P_\perp a\|$$

Which means that the suboptimality of the obtained norm $\leq \|P_\perp a\|$

as required.

□