

Foundations of Deep Learning – Homework Assignment #4

Adi Alm & Tomer Epshtain

Part 1:

Experiment:

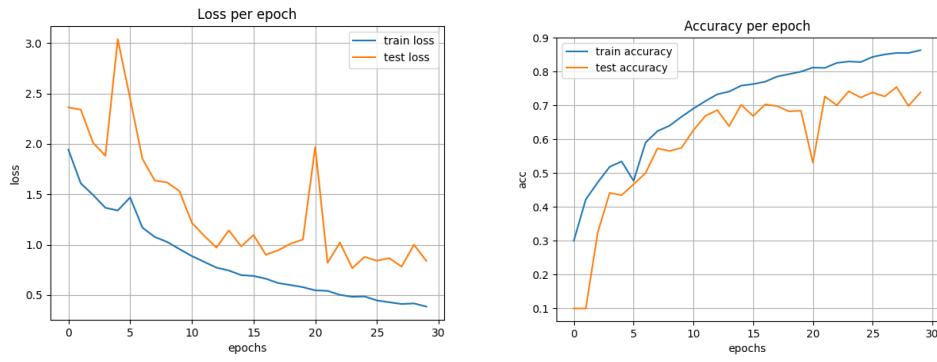
Demonstrate the four empirical postulates we relied on in class when rationalizing about generalization in deep learning.

Results:

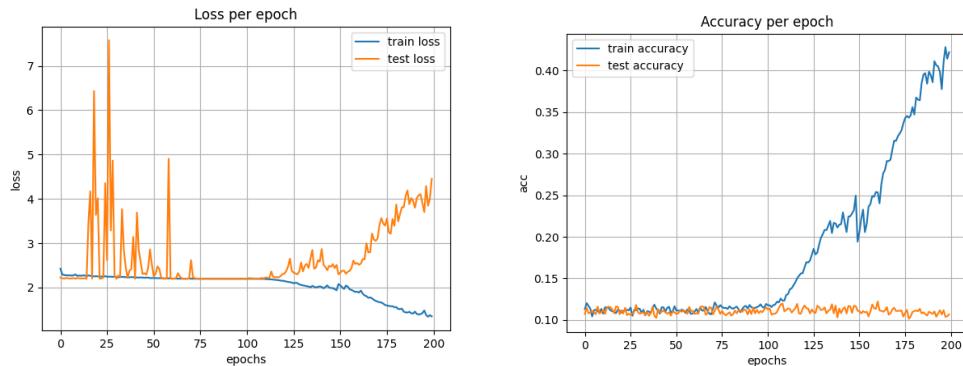
We chose to work with the CIFAR10 dataset with the common MobileNetV2 architecture.

Here are our results:

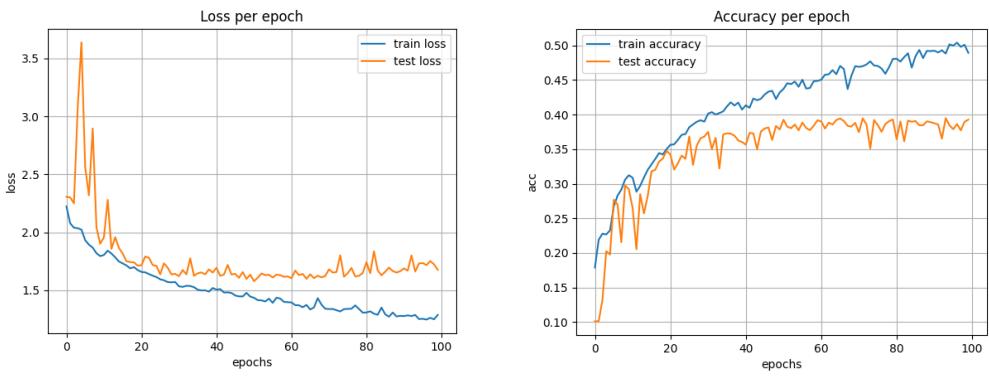
1. Standard NNs (=MobileNetV2) trained by a standard optimization algorithm (=Adam) on a standard dataset (=CIFAR10) generalize well without any explicit regularization. Even though # of learned parameters \gg # of training examples.



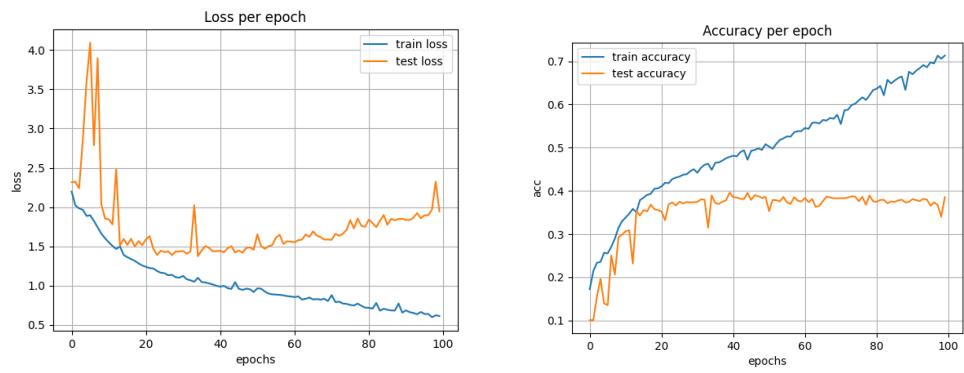
2. In the regime of (1.), the training error of learned hypothesis ≈ 0 . This remains the case even if we replace the training examples by any training set of the same size. We demonstrated this by replacing CIFAR10 dataset with random images (with every pixel randomly samples from a uniform 0,1 distribution) 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 ≈ 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

$$H = \left\{ \mathbb{R}^d \ni x \mapsto y = W_0 \sigma(W_1 \sigma(W_2 \sigma(\dots W_{N-1} \sigma(W_N x) \dots)) \in \mathbb{R}^d : W_n \in \mathbb{R}^{d \times d}, n \in [N] \right\}$$

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

For $r \in \mathbb{N}$ 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.:

$$H_r = \{x \mapsto y = W_N V_N^T \sigma(W_{N-1} V_{N-1}^T \dots \sigma(W_1 V_1^T x) \dots) : W_n, V_n \in \mathbb{R}^{d \times r}, n \in [N]\}$$

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

Given a loss $\ell: \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, 1]$ s.t. $\forall y, y' | \ell(y, \hat{y}) - \ell(y, y') | \leq \beta \|y - y'\|$

We would like to derive generalization bounds for H .

(a) Fix $r \in \mathbb{N}$ 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 $r \in \mathbb{N}$

Solution (a)

Let $r \in \mathbb{N}$.

By assumption each of the $2Ndr$ parameters representing H_r using

$$\text{bits, thus } |H_r| \leq \underbrace{2^b \cdots 2^b}_{\substack{\# \text{options} \\ \text{for the} \\ \text{first param}}} = 2^{\substack{b \cdot 2Ndr \\ \# \text{options} \\ \text{for the} \\ 2Ndr \text{ param}}}$$

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

For \hat{h} the returned hypothesis from the algorithm Adecon w.p. $\geq 1 - \delta$ over SMDM:

$$L_0(\hat{h}) - L_S(\hat{h}) \leq \sqrt{\frac{(b \cdot 2Ndr + 1) \ln(2) + \ln(\frac{1}{\delta})}{2m}} + 2 \cdot \rho \cdot d(\hat{h}, H_r)$$

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):

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

For \hat{h} with weighted matrices W_N, \dots, W_1

and h_r with weighted matrices W_N', \dots, W_1' with rank $\leq r$
the closest approximations to W_N, \dots, W_1 .

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

Note, that this equation holds for approximations to rank $r \geq 1$ too,

since if we denote W_n as spectral decomposition $W_n = \sum_{i=1}^d \sigma_i u_i v_i^\top$

where $|\sigma_1| \geq |\sigma_2| \geq \dots \geq |\sigma_d|$ then the best rank- r approximation

to W_n is $W_n' = \sum_{i=1}^r \sigma_i u_i v_i^\top$, thus we can see clearly

that $\|W_n'\|_{\text{spectral}} = |\sigma_1| = \|W_n\|_{\text{spectral}}$ and thus the

bound denoted above holds for each $n \in [d]$.

$$\begin{aligned} \text{In addition } \|W_n - W_n'\|_{\text{spectral}} &= \left\| \sum_{i=1}^d \sigma_i u_i v_i^\top - \sum_{i=1}^r \sigma_i u_i v_i^\top \right\|_{\text{F}} = \left\| \sum_{i=r+1}^d \sigma_i u_i v_i^\top \right\|_{\text{F}} = |\sigma_{r+1}| \\ &= |\sigma_{r+1}(W_n)| \end{aligned}$$

\Rightarrow In total we'll get $\sqrt{\epsilon_{\text{recom}}}$ w.p. $\geq 1 - \delta$ over $S \sim D^m$:

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

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 $S^M D^m$ and fixed $1 \leq r \leq d$:

$$\begin{aligned} L_D(\hat{h}) - L_S(\hat{h}) &\leq \sqrt{\frac{(b \cdot 2Nd(r+1))\ln(2) + \ln(\frac{1}{\delta})}{2(M-1)}} + 2\beta \cdot \left[\delta^{N-1} \sum_{n=1}^N \prod_{j \neq n} \|W_j\|_{sp} \cdot |\sigma_{rn}(W_n)| \right] \\ &= \sqrt{\frac{(b \cdot 2Nd(r+1))\ln(2) + \ln(\frac{d}{\delta})}{2(M-1)}} + 2\beta \cdot \left[\delta^{N-1} \sum_{n=1}^N \prod_{j \neq n} \|W_j\|_{sp} \cdot |\sigma_{rn}(W_n)| \right] \end{aligned}$$

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

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

$$\Rightarrow \Pr(\exists r L_D(\hat{h}) - L_S(\hat{h}) \geq A_r + B_r) \leq \sum_{r=1}^d \Pr(L_D(\hat{h}) - L_S(\hat{h}) \geq A_r + B_r) \leq \sum_{r=1}^d \frac{\delta}{d} = \delta$$

union bound

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

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

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

□

Foundations of Deep Learning – Homework Assignment #4

Adi Almog & Tomer Epshteyn

Part 2: (2)

Question:

Let $H = \{h_\theta: \mathcal{X} \rightarrow \mathcal{Y} : \theta \in \mathbb{R}^p, \|\theta\|_\infty \leq 0.5\}$ 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 $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow [0,1]$ and a training set $S = \{(x_i, y_i)\}_{i=1}^m \subseteq \mathcal{X} \times \mathcal{Y}$, the Radamacher complexity of \mathcal{H}_Θ 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:

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

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:

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

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

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

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

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

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

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

$\forall a > 0$:

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

We want to choose a such that (?) holds

i.e.

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

Choose:

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

So we have

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

Or by viewing complement:

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

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:

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

- On the one hand:

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

- On the other hand:

$$\begin{aligned} & \Pr\left(\left(\forall h \in \mathcal{H}_k: L_D(h) - L_S(h) \leq 2R(\ell \circ \mathcal{H}_k \circ S) + \sqrt{\frac{2 \cdot \ln\left(\frac{2\pi^2}{3} k^2 \frac{2}{\delta}\right)}{m}}\right) \cap \left(R(\ell \circ \mathcal{H}_{\Theta_k} \circ S) < \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/2}\right)\right) \leq \\ & \leq \Pr\left(\forall h \in \mathcal{H}_k: L_D(h) - L_S(h) < \frac{\left(\frac{1}{2}\right)^p - \left(\frac{1}{2} - k\epsilon\right)^p}{\delta/4} + \sqrt{\frac{2 \cdot \ln\left(\frac{2\pi^2}{3} k^2 \frac{2}{\delta}\right)}{m}}\right) \end{aligned}$$

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 Σ_0 and Σ_1 are non-singular (positive definite). Then it holds that:

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

Solution

First, let's sign the two distributions as:

$$P = N(\mu_0, \Sigma_0) \quad \text{and} \quad Q = N(\mu_1, \Sigma_1).$$

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

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

$$\begin{aligned} \frac{p(x)}{q(x)} &= p(x) \cdot \frac{1}{q(x)} = \frac{1}{(2\pi)^{\frac{r}{2}} (\det \Sigma_0)^{\frac{1}{2}}} \cdot e^{-\frac{1}{2} (x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0)} \cdot \frac{(2\pi)^{\frac{r}{2}} (\det \Sigma_1)^{\frac{1}{2}}}{e^{-\frac{1}{2} (x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)}} \\ &= \frac{(\det \Sigma_1)^{\frac{1}{2}}}{(\det \Sigma_0)^{\frac{1}{2}}} \cdot e^{-\frac{1}{2} (x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0) - (-\frac{1}{2} (x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1))} \\ &= \frac{(\det \Sigma_1)^{\frac{1}{2}}}{(\det \Sigma_0)^{\frac{1}{2}}} \cdot e^{\frac{1}{2} \cdot [(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) - (x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0)]} \end{aligned}$$

Thus:

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

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

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

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

linearity of $\mathbb{E}(\cdot)$ \rightarrow

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

$$- \mathbb{E}_P \left[\frac{1}{2} (x - \mu_0)^T \Sigma_0^{-1} (x - \mu_0) \right]$$

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

Claim

For each $x \in \mathbb{R}^r, A \in \mathbb{R}^{r \times r}$: $x^T A x = \text{tr}(A x x^T)$

Proof

$$x^T A x = (x_1 \cdots x_r) \begin{pmatrix} 1 \\ A \\ x \end{pmatrix} = (x_1 \cdots x_r) \begin{pmatrix} (Ax)_1 \\ \vdots \\ (Ax)_r \end{pmatrix} = (x_1 \cdots x_r) \begin{pmatrix} \sum_{j=1}^r a_{1j} x_j \\ \vdots \\ \sum_{j=1}^r a_{rj} x_j \end{pmatrix} =$$

$$= \sum_{j=1}^r a_{1j} x_1 x_j + \cdots + \sum_{j=1}^r a_{rj} x_r x_j =$$

$$= \sum_{j=1}^r a_{1j} (x x^T)_{1j} + \cdots + \sum_{j=1}^r a_{rj} (x x^T)_{rj} =$$

$$= (A x x^T)_{11} + \cdots + (A x x^T)_{rr} = \text{tr}(A x x^T)$$

□

Continue next page...

Thus:

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

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

$$\begin{aligned} & \text{tr}(E[A]) = E[\text{tr}(A)] \\ \text{by linearity} & \rightarrow = \frac{1}{2} \left[\ln \left(\frac{\det(\Sigma_1)}{\det(\Sigma_0)} \right) + \text{tr}(\mathbb{E}_P[\Sigma_1^{-1} (x - \mu_1)(x - \mu_1)^T]) - \text{tr}(\mathbb{E}_P[\Sigma_0^{-1} (x - \mu_0)(x - \mu_0)^T]) \right] \\ & = \frac{1}{2} \left[\ln \left(\frac{\det(\Sigma_1)}{\det(\Sigma_0)} \right) + \text{tr}(\Sigma_1^{-1} \mathbb{E}_P[x x^T - \mu_1 x^T - x \mu_1^T + \mu_1 \mu_1^T]) \right. \\ & \quad \left. - \text{tr}(\Sigma_0^{-1} \mathbb{E}_Px - \mu_0^T) \right] \quad \text{②} \end{aligned}$$

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

$$\mathbb{E}_P[x] = \mu_0 \text{ and } \mathbb{E}_P[(x - \mu_0)(x - \mu_0)^T] = \Sigma_0$$

$$\text{and } \mathbb{E}_P[x x^T] = \Sigma_0 + \mu_0 \mu_0^T, \text{ so:}$$

$$\begin{aligned} \text{②} & \frac{1}{2} \left[\ln \left(\frac{\det(\Sigma_1)}{\det(\Sigma_0)} \right) + \text{tr}(\Sigma_1^{-1} (\Sigma_0 + \mu_0 \mu_0^T - \mu_1 \mu_0^T - \mu_0 \mu_1^T + \mu_1 \mu_1^T)) \right. \\ & \quad \left. - \text{tr}(\Sigma_0^{-1} \Sigma_0) \right] \end{aligned}$$

$$\begin{aligned} & = \frac{1}{2} \left[\ln \left(\frac{\det(\Sigma_1)}{\det(\Sigma_0)} \right) + \text{tr}(\Sigma_1^{-1} \Sigma_0) + \text{tr}(\Sigma_1^{-1} (\mu_0 \mu_0^T - \mu_1 \mu_0^T - \mu_0 \mu_1^T + \mu_1 \mu_1^T)) \right. \\ & \quad \left. - \text{tr}(\Sigma_0) \right] \end{aligned}$$

$$\begin{aligned} & X^T A X \\ & = A X X^T \\ & \text{by claim} \\ & \text{we prove!} \end{aligned} \quad \leftarrow = \frac{1}{2} \left[\ln \left(\frac{\det(\Sigma_1)}{\det(\Sigma_0)} \right) - r + \text{tr}(\Sigma_1^{-1} \Sigma_0) + \text{tr}(\mu_0^T \Sigma_1^{-1} \mu_0 - \mu_0^T \Sigma_1^{-1} \mu_1 - \mu_1^T \Sigma_1^{-1} \mu_0 + \mu_1^T \Sigma_1^{-1} \mu_1) \right]$$

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

\Rightarrow

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

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 least one of a finite set of points $\{\theta_1, \dots, \theta_K\}$.

Using PAC-BAYES derive a generalization bound which accounts for the latent implicit regularization.

Solution

For $1 \leq i \leq K$ 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:

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

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. $\geq 1 - \frac{\delta}{K}$ over $S \sim D^M$:

$$L_D(Q) - L_S(Q) \leq \sqrt{\frac{KL(Q || P_i) + \ln(\frac{2M}{\delta K})}{2(M-1)}} = \sqrt{\frac{\frac{1}{2\sigma^2} \|\hat{\theta} - \theta_i\|^2 + \ln(\frac{2M}{\delta K})}{2(M-1)}}$$

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

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

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

$$\Rightarrow \Pr(L_D(Q) - L_S(Q) \geq \sqrt{\frac{\frac{1}{2\sigma^2} \|(\hat{\theta} - \theta_i)\|^2 + \ln(\frac{2mK}{\delta})}{2(m-1)}}) \leq \frac{\delta}{K}$$

$$\Rightarrow \Pr(\exists i : L_D(Q) - L_S(Q) \geq \sqrt{\frac{\frac{1}{2\sigma^2} \|(\hat{\theta} - \theta_i)\|^2 + \ln(\frac{2mK}{\delta})}{2(m-1)}}) \quad (2)$$

$$\begin{aligned} \text{Union bound } \rightarrow (2) & \sum_{i=1}^K \Pr(L_D(Q) - L_S(Q) \geq \sqrt{\frac{\frac{1}{2\sigma^2} \|(\hat{\theta} - \theta_i)\|^2 + \ln(\frac{2mK}{\delta})}{2(m-1)}}) \\ & \leq \sum_{i=1}^K \frac{\delta}{K} = \delta \end{aligned}$$

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

$$\begin{aligned} L_D(Q) - L_S(Q) & \leq \min_{1 \leq i \leq K} \sqrt{\frac{\frac{1}{2\sigma^2} \|(\hat{\theta} - \theta_i)\|^2 + \ln(\frac{2mK}{\delta})}{2(m-1)}} \\ & = \sqrt{\frac{\frac{1}{2\sigma^2} \cdot \min_{1 \leq i \leq K} \|(\hat{\theta} - \theta_i)\|^2 + \ln(\frac{2mK}{\delta})}{2(m-1)}} \end{aligned}$$

Since $\hat{\theta}$ tends to ~~not~~ be close to at least one of $\{\theta_i\}_{i=1}^K$

we'll get that $\min_{1 \leq i \leq K} \|(\hat{\theta} - \theta_i)\|^2$ will be with low value.

In addition by assumption the optimization tends to ~~not~~ have 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.

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

□

Learned with Ca

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 $w^{(0)} = \alpha \in \mathbb{R}^d$ and producing iterates $w^{(1)}, w^{(2)}, \dots$ via iterative algorithm in which every update $w^{(t+1)} - w^{(t)} \in \text{span}\{\nabla L_{(x_i, y_i)}(w) : i \in [m], w \in \mathbb{R}^d\}$.

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 $\leq \|P_\perp \alpha\|$ where $P_\perp : \mathbb{R}^d \rightarrow \mathbb{R}^d$ stands for projection onto the orthogonal complement of $\text{span}\{x_i\}_{i=1}^m$.

Proof

Denote by $w^* = \arg \min_{\substack{w \in \mathbb{R}^d \\ \text{global min} \\ L_s(w)=0}} \|w\|$ the solution

with the min norm across all global optima.

Lemma 1

The solution under the settings in the proposition is

$$\tilde{w} = \alpha + X(X^T X)^{-1}(Y - X^T \alpha)$$

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

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

in order to finish the question.

Proof of lemma 1

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

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

Step: Suppose $w^t \in \alpha + \text{span}\{x_i\}_{i=1}^m$. We'll prove that $w^{t+1} \in \alpha + \text{span}\{x_i\}_{i=1}^m$
 $w^t \in \alpha + \text{span}\{x_i\}_{i=1}^m \Rightarrow w^t = \alpha + \sum_{i=1}^m \alpha_i x_i$ for $\{\alpha_i\}_{i=1}^m$ scalars.

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

$$\nabla_{x_i} (w \cdot x_i, y_i) = (x_i^\top w - y_i) \cdot x_i$$

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 $w^{t+1} - w^t \in \text{span}\{\nabla_{x_i} (w) | i \in [m], w \in \mathbb{R}^d\}$)

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

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

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

□

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

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

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

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

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

(continue of the proof next page)

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

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

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

$$\Rightarrow X^T\alpha + X^T X r = y$$

$$\Rightarrow X^T X r = y - X^T \alpha$$

$$\text{rank}(X^T X) = n \Rightarrow r = (X^T X)^{-1} (y - X^T \alpha)$$

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

as required. \square

In class

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

Thus we have to show that

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

We'll show that:

$$\| \alpha + X(X^T X)^{-1} (y - X^T \alpha) \| = \| \alpha + X(X^T X)^{-1} y - X(X^T X)^{-1} X^T \alpha \|$$

$$\leq \| X(X^T X)^{-1} y \| + \| \alpha - X(X^T X)^{-1} X^T \alpha \|$$

▫ inequality

$$\leq \| X(X^T X)^{-1} y \| + \| P_{\perp} \alpha + P_{\parallel} \alpha - X(X^T X)^{-1} X^T (P_{\perp} \alpha + P_{\parallel} \alpha) \|$$

$$\alpha = P_{\parallel} \alpha + P_{\perp} \alpha$$

where $P_{\parallel} \alpha \in \text{span}\{x_i\}_{i=1}^m$ and $P_{\perp} \alpha$ is the projection as described in the question.

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

thus:

$$\Leftrightarrow \| X(X^T X)^{-1} y \| + \| P_{\perp} \alpha + P_{\parallel} \alpha - X(X^T X)^{-1} X^T P_{\parallel} \alpha \| \Leftrightarrow$$

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

$$\Leftrightarrow \| X(X^T X)^{-1} y \| + \| P_{\perp} \alpha + Xb - X(X^T X)^{-1} X^T Xb \|$$

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

Which means that the suboptimality of the obtained norm $\leq \| P_{\perp} \alpha \|$

as required. \square

Matrix factorization (bonus)

Let $W^* \in \mathbb{R}^{d \times d}$ s.t. $W^* = U \Sigma V^T$ SVD of W^*
 $U \in \mathbb{R}^{d \times d}, \Sigma \in \mathbb{R}^{d \times d}, V \in \mathbb{R}^{d \times d}$

Let the training loss be $L_s(W) = \frac{1}{d \cdot d'} \cdot \frac{1}{2} \|W - W^*\|_{Fro}^2$

Suppose we optimize $L_s(\cdot)$ by overparameterization with LNN
of depth N and gradient flow starting from balanced init.

Suppose we can write $W_{1:N}(t) = U(t) \Sigma(t) V^T$

For depth $N=1, 2$ derive explicit expressions for $\{\sigma_r(t)\}_{r=1}^{\min\{d, d'\}}$

singular values of $W_{1:N}(t)$.

Explain how added depth ($N=2$) leads singular values to

"shoot up" one by one, in line with the interpretation given in class.

Solution

Remember the theorem we proved in class!

Assume we run GF over ϕ starting from balanced init.

Then the singular values of $W_{1:N}(t)$ evolve by:

$$\dot{\sigma}_r(t) := \frac{d}{dt} \sigma_r(t) = -N \cdot (\sigma_r(t)^2)^{1-\frac{1}{N}} \langle \nabla L_s(W_{1:N}(t)), U_r(t) V_r(t)^T \rangle$$

$r \in [\min\{d, d'\}]$ for $W_{1:N}(t) = U(t) \Sigma(t) V^T$

In this case $U(t) = U$, $V(t) = V$ are fixed,

Thus:

$$\begin{aligned} \text{For } N=1: \quad \dot{\sigma}_r(t) &= -1 \cdot (\sigma_r(t)^2)^{1-\frac{1}{N}} \langle \nabla L_s(W_{1:N}(t)), U_r V_r^T \rangle \\ &= - \langle \nabla L_s(W_1(t)), U_r V_r^T \rangle \end{aligned}$$

$$\begin{aligned} \text{For } N=2: \quad \dot{\sigma}_r(t) &= -2 \cdot (\sigma_r(t)^2)^{1-\frac{1}{N}} \langle \nabla L_s(W_{1:N}(t)), U_r V_r^T \rangle \\ &= -2 |\sigma_r(t)| \langle \nabla L_s(W_{1:2}(t)), U_r V_r^T \rangle \end{aligned}$$

$$\text{Note: } L(w) = \frac{1}{d \cdot d'} \cdot \frac{1}{2} \|w - w^*\|_F^2$$

$$\frac{\partial L}{\partial w_{ij}}(w) = \frac{1}{d \cdot d'} \cdot \frac{1}{2} \frac{\partial}{\partial w_{ij}} (w_{ij} - w_{ij}^*)^2 = \frac{1}{d \cdot d'} (w_{ij} - w_{ij}^*)$$

$$\nabla L(w) = \frac{1}{d \cdot d'} (w - w^*)$$

\Rightarrow For both $N=1, 2$ it holds:

$$\langle \nabla L(w_{1:N}(t)), U_r V_r^\top \rangle = \cancel{\frac{1}{d \cdot d'} \langle w - w^*, U_r V_r^\top \rangle}$$

$$\text{for } w^* = \cancel{U} \Sigma V^\top \rightarrow = \frac{1}{d \cdot d'} \langle w_{1:N}(t) - w^*, U_r V_r^\top \rangle$$

$$= \frac{1}{d \cdot d'} \langle U(S(t) - \Sigma)V^\top - U\Sigma V^\top, U_r V_r^\top \rangle$$

$$= \frac{1}{d \cdot d'} \langle U(S(t) - \Sigma)V^\top, U_r V_r^\top \rangle$$

$$\text{tr}(\cdot) \text{ is cyclic} \quad = \frac{1}{d \cdot d'} \cancel{\text{tr}}(V(S(t) - \Sigma)U^\top U_r V_r^\top)$$

$$R = \frac{1}{d \cdot d'} \text{tr}(V_r^\top V(S(t) - \Sigma)U^\top U_r)$$

$$= \frac{1}{d \cdot d'} \text{tr}(e_r^\top (S(t) - \Sigma) e_r)$$

$$= \frac{1}{d \cdot d'} \cdot \widehat{o_r(t)} - \widehat{o_f}$$

$$\left[\Sigma = \begin{pmatrix} \sigma_1 & \dots & \sigma_n \end{pmatrix} \right] \rightarrow$$

Continue next page...

Now, we'll solve the general equation:

$$\dot{x}(t) = a \cdot x(t) + b \quad \text{for fixed } a, b.$$

||

$$\frac{\dot{x}(t)}{ax(t)+b} = 1$$

||

$$\frac{1}{a} \cdot \frac{d}{dt} [\ln(ax(t)+b)] = 1$$

||

$$\frac{d}{dt} \ln(ax(t)+b) = a$$

||

$$\int_0^t \frac{d}{dt} \ln(ax(\epsilon)+b) d\epsilon = \int_0^t a + d$$

$$\ln(ax(t)+b) - \ln(ax(0)+b) = a \cdot t + d$$

$$\ln(ax(t)+b) = a \cdot t + \underbrace{d + \ln(ax(0)+b)}_{\substack{\text{we'll sign as } c \\ \text{and we can find it}}} = a \cdot t + c$$

e^t

$$\Rightarrow ax(t)+b = e^{at+c} = e^{at} \cdot e^c$$

$$\Rightarrow x(t) = \frac{e^{at} \cdot e^c - b}{a}$$

in case we want $x(0) = \varepsilon$ we'll get:

$$\varepsilon = x(0) = \frac{e^c - b}{a} \Rightarrow a \cdot \varepsilon = e^c - b$$

$$\Rightarrow c = \ln(a\varepsilon + b)$$

$$\Rightarrow x(t) = \frac{e^{at} \cdot e^{\ln(a\varepsilon+b)} - b}{a} = \frac{(a\varepsilon+b) \cdot e^{at} - b}{a}$$

$$\Rightarrow \text{in our case } \sigma_r(t) = - \frac{\sigma_r(t)}{d \cdot d'} + \cancel{\frac{(\varepsilon - \sigma_r(t))}{d \cdot d'}} \frac{\sigma_r}{d \cdot d'}$$

$$\Rightarrow a = -\frac{1}{d \cdot d'} \quad b = \cancel{\frac{(\varepsilon - \sigma_r(t))}{d \cdot d'}} \frac{\sigma_r}{d \cdot d'}$$

For $N=1$:

$$\sigma_r(t) = \frac{\left(\frac{-\varepsilon}{d \cdot d'} + \frac{\sigma_r}{d \cdot d'} \right) \cdot e^{-\frac{1}{d \cdot d'} t} - \frac{\sigma_r}{d \cdot d'}}{-\frac{1}{d \cdot d'}} \quad (\cancel{(-\varepsilon + \sigma_r)e^{-\frac{1}{d \cdot d'} t}})$$
$$= (-\sigma_r + \varepsilon) e^{-\frac{1}{d \cdot d'} t} + \sigma_r$$

For $N=2$ we've got:

$$\begin{aligned}\dot{\sigma}_r(t) &= -2\sigma_r(t) \cdot \frac{1}{d \cdot d'} (\sigma_r(t) - \sigma_r) \\ &= -\frac{2}{d \cdot d'} \sigma_r(t)^2 + \frac{2}{d \cdot d'} \sigma_r(t)\end{aligned}\quad (*)$$

which of the form $\dot{x}(t) = ax(t)^2 + bx(t)$

the solution will be

$$x(t) = \frac{b}{e^{-bt-c} - a}$$

By requiring : $x(0) = \varepsilon$ we get

$$\varepsilon = x(0) = \frac{b}{e^{-c-a}}$$

$$\varepsilon \cdot e^c - a\varepsilon = b$$

$$\varepsilon \cdot e^c = b + a\varepsilon$$

$$e^c = \frac{b + a\varepsilon}{\varepsilon}$$

$$-c = \ln\left(\frac{b + a\varepsilon}{\varepsilon}\right)$$

$$c = -\ln\left(\frac{b + a\varepsilon}{\varepsilon}\right)$$

$$\Rightarrow x(t) = \frac{b}{e^{-bt+\ln\left(\frac{b+a\varepsilon}{\varepsilon}\right)} - a}$$

$$\Rightarrow \text{For } N=2 \quad (*) \quad a = \frac{-2}{d \cdot d'} \quad b = \frac{2\sigma_r}{d \cdot d'}$$

we'll see:

$$\begin{aligned}\sigma_r(t) &= \frac{\frac{2\sigma_r}{d \cdot d'}}{e^{-\frac{2\sigma_r t}{d \cdot d'} + \ln\left(\frac{-2 \cdot \varepsilon + 2\sigma_r}{d \cdot d'}\right)} + \frac{2}{d \cdot d'}} = \frac{\frac{2\sigma_r}{d \cdot d'}}{\left(\frac{-2 \cdot \varepsilon + 2\sigma_r}{d \cdot d'}\right) \cdot e^{-\frac{2\sigma_r t}{d \cdot d'}} + \frac{2}{d \cdot d'}}\end{aligned}$$

$$\text{For } N=1 \quad \sigma_r(t) = (-\sigma_r + \varepsilon) \cdot e^{-\frac{1}{d \cdot d'} t} + \sigma_r$$

As we can clearly see, in the exponent in the case $N=2$ →

we have $e^{-\frac{2\sigma_r t}{d \cdot d'}}$ and in $N=1$ $e^{-\frac{1}{d \cdot d' t}}$, thus for large singular values σ_r we'll converge faster to 0 in the second case and thus in total converge faster to σ_r .

