**Machine Learning** **Fall 2023**

# Lecture 11: Dimensionality Reduction, Algorithm Stability, von Neumann's Minmax Theorem

*Lecturer: Liwei Wang*

*Scribe: Yichao Zhou, Meng Yuan, Ruihuan Wang, Xinjie Gao, Ruichen Wang, Jiahe Geng*



## 11.1 Dimensionality Reduction

### 11.1.1 Review on PCA

Assuming we have n points $x_1, x_2, ...x_n \in \mathbb{R}^d$, we want to find a linear mapping $\Phi : \mathbb{R}^d \to \mathbb{R}^{d'} (d' < d)$, such that the distance between the original point and the mapped point is approximately equal.

$$\begin{aligned} \min_{\Phi} \quad & \sum_{i=1}^{n} ||x_i - \Phi(x_i)||^2 \\ s.t. \quad & \Phi : \mathbb{R}^d \to \mathbb{R}^{d'} (d' < d) \end{aligned}$$

### 11.1.2 JL Lemma

Instead of minimizing the projection distance, we can consider the distance between any pair of data points. We hope to find a linear mapping that hardly changes the distance between any pair of data points. We can formulate the problem as follows.

$$(1-\epsilon)\|\mathbf{x_i} - \mathbf{x_j}\|^2 \leq \|\Phi(\mathbf{x_i}) - \Phi(\mathbf{x_j})\|^2 \leq (1+\epsilon)\|\mathbf{x_i} - \mathbf{x_j}\|^2, \ \forall i, j \in [n], \tag{11.1}$$

where $x_i, i \in [n]$ and $\Phi$ have the same definition as 11.1.1, and $\epsilon > 0$ is a given constant.

JL Lemma shows that, if $d' \geq \frac{8 \ln n}{\epsilon^2}$, then there exists such a linear mapping. To make the proof, we begin with 3 lemmas.

**Lemma 11.1 (The moment-generating function of $\chi^2$ random variables)** *Let $Q$ be a ramdom variable following a $\chi^2$ distribution with $k$ degrees of freedom. Its moment-generating function*

$$\mathbb{E}[e^{tQ}] = (1-2t)^{-\frac{k}{2}}, \ t \in (0, \frac{1}{2}).$$

**Proof:** Since Q is a ramdom variable following a $\chi^2$ distribution with $k$ degrees of freedom, we have

$$Q = \sum_{i=1}^{k} X_i^2, \ X_i \underset{i.i.d}{\sim} N(0,1), \ \forall i \in [k].$$

Then,

$$\mathbb{E}[e^{tQ}] = (\mathbb{E}[e^{tX_1}])^k.$$

And then we can calculate $\mathbb{E}[e^{tX_1}]$ by

$$\begin{aligned}\mathbb{E}[e^{tX_1}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{tx^2} e^{-\frac{x^2}{2}} \, \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi(1-2t)}} \int_{-\infty}^{+\infty} e^{-\frac{(\sqrt{1-2t}x)^2}{2}} \, \mathrm{d}\sqrt{1-2t}x \\ &= (1-2t)^{-\frac{1}{2}}.\end{aligned}$$

From which we can get

$$\mathbb{E}[e^{tQ}] = (1-2t)^{-\frac{k}{2}}.$$

■

**Lemma 11.2** *Let $Q$ be a ramdom variable following a $\chi^2$ distribution with $k$ degrees of freedom, then*

$$\mathbb{P}[(1-\epsilon)k \leq Q \leq (1+\epsilon)k] \geq 1 - 2e^{-(\epsilon^2-\epsilon^3)k/4}, \forall \epsilon \in (0, \frac{1}{2}).$$

**Proof:** Consider one side,

$$\mathbb{P}(Q \geq (1+\epsilon)k) = \mathbb{P}(e^{tQ} \geq e^{(1+\epsilon)tk}), \forall t \in (0, \frac{1}{2}).$$

Basically, we hope to achieve a better bound utilizing the moment-generating function. Firstly, we can apply Markov's Inequality.

$$\mathbb{P}[e^{tQ} \geq e^{(1+\epsilon)tk}] \leq \frac{\mathbb{E}[e^{tQ}]}{e^{(1+\epsilon)tk}} = \frac{(1-2t)^{-\frac{k}{2}}}{e^{(1+\epsilon)tk}}.$$

By taking the derivative, the minimal point of the right side is $t = \frac{\epsilon}{2(1+\epsilon)}$, so

$$\mathbb{P}(Q \geq (1+\epsilon)k) \leq (\frac{1+\epsilon}{e^{\epsilon}})^{\frac{k}{2}}.$$

Using Taylor's formula for $\ln(1+x)$, we can get

$$\ln(1+\epsilon) \leq \epsilon - \frac{\epsilon^2}{2} + \frac{\epsilon^3}{3} \leq \epsilon - \frac{\epsilon^2 - \epsilon^3}{2}, \epsilon \in (0, \frac{1}{2}).$$

Thus,

$$\begin{aligned}1 + \epsilon &\leq e^{\epsilon} \cdot e^{-\frac{\epsilon^2-\epsilon^3}{2}}, \\ \frac{1+\epsilon}{e^{\epsilon}} &\leq e^{-\frac{\epsilon^2-\epsilon^3}{2}}.\end{aligned}$$

So we get

$$\mathbb{P}(Q \geq (1+\epsilon)k) \leq (\frac{1+\epsilon}{e^{\epsilon}})^{\frac{k}{2}} \leq e^{-(\epsilon^2-\epsilon^3)k/4}$$

The same goes with the other side, and by applying the union bound, we reach

$$\mathbb{P}[(1-\epsilon)k \leq Q \leq (1+\epsilon)k] \geq 1 - 2e^{-(\epsilon^2-\epsilon^3)k/4}, \forall \epsilon \in (0, \frac{1}{2}).$$

■

Based on 11.2, we introduce the following lemma.

**Lemma 11.3** *For $\mathbf{x} \in \mathbb{R}^N$, we define $k < N$ and assume the elements from the matrix $\mathbf{A} \in \mathbb{R}^{k\times N}$ are independently sampled from normal distribution $N(0,1)$. For any $\epsilon \in (0, \frac{1}{2})$,*

$$\mathbb{P}[(1-\epsilon)\|\mathbf{x}\|^2 \leq \|\frac{1}{\sqrt{k}}\mathbf{A}\mathbf{x}\|^2 \leq (1+\epsilon)\|\mathbf{x}\|^2] \geq 1 - 2e^{-(\epsilon^2-\epsilon^3)k/4}.$$

**Proof:** Let $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_k)^T = \mathbf{A}\mathbf{x} \in \mathbb{R}^k$.

Since $\hat{x}_i,\ \forall i \in [k]$ are linear combinations of gaussian variables $A_{ij}, j \in [N]$ with zero expectations, they are gaussian variables with zero expectations as well. What's more,

$$\mathbb{E}[\hat{x}_i^2] = \mathbb{E}[(\sum_{j=1}^{N} A_{ij}x_j)^2].$$

As $A_{ij}$ are sampled independently, all the cross terms

$$\mathbb{E}[A_{ij_1}A_{ij_2}x_{j_1}x_{j_2}] = x_{j_1}x_{j_2}\mathbb{E}[A_{ij_1}]\mathbb{E}[A_{ij_2}] = 0, \forall i \in [k], j_1, j_2 \in [N],$$

and as the variance of the normal distribution is 1, the second-order moment for all $A_{ij}$ are 1, thus

$$\mathbb{E}[A_{ij}^2 x_j^2] = \mathbb{E}[A_{ij}^2]x_j^2 = x_j^2, \forall i \in [k], j \in [N].$$

Hence,

$$\text{var}[\hat{x}_i^2] = \mathbb{E}[\hat{x}_i^2] = \mathbb{E}[(\sum_{j=1}^{N} A_{ij}x_j)^2] = \sum_{j=1}^{N}\mathbb{E}[A_{ij}^2x_j^2] = \sum_{j=1}^{N} x_j^2 = \|\mathbf{x}\|^2, \forall i \in [k].$$

Let $T_i = \frac{\hat{x}_i}{\|x\|}, \forall i \in [k]$, we know $T_i \sim N(0,1)$ and are independent. Thus, $\sum_{i=1}^{k} T_i^2 = \frac{\sum_{i=1}^{k} \hat{x_i}^2}{\|x\|^2} = \frac{\|\hat{\mathbf{x}}\|^2}{\|\mathbf{x}\|^2}$ are $\chi^2$ variables with $k$ degrees of freedom. Using this and 11.2, the inequality of this lemma is trivial. ■

Finally we can prove JL Lemma.

**Lemma 11.4 (Johnson-Lindenstrauss)** *Given data points $x_1, x_2, ...x_n \in \mathbb{R}^d$, $\epsilon \in (0, \frac{1}{2})$, if $d' \geq \frac{8\ln n}{\epsilon^2}$, then there exists a linear mapping $\Phi : \mathbb{R}^d \to \mathbb{R}^{d'}$, such that $\forall i, j \in [n]$, Equation (11.1) holds true.*

**Proof:** We choose $f = \frac{1}{\sqrt{d'}}\mathbf{A}$, where $\mathbf{A} \in \mathbb{R}^{d'\times d}$, and all the elements are independently sampled from $N(0,1)$. According to 11.3, $\forall i, j \in [n]$ we have

$$\mathbb{P}[(1-\epsilon)\|\mathbf{x_i} - \mathbf{x_j}\|^2 \leq \|\frac{1}{\sqrt{d'}}(\Phi(\mathbf{x_i}) - \Phi(\mathbf{x_j}))\|^2 \leq (1+\epsilon)\|\mathbf{x_i} - \mathbf{x_j}\|^2] \geq 1 - 2e^{-(\epsilon^2-\epsilon^3)d'/4}.$$

This can be considered as a success rate. In contrast, we consider the fail rate, i.e., for any given pair $(i,j), i, j \in [n]$,

$$\mathbb{P}[fail] \leq 2e^{-(\epsilon^2-\epsilon^3)d'/4}.$$

Applying the union bound, for all the $\frac{n(n-1)}{2}$ pairs of data, we hope

$$\mathbb{P}[totalfail] \leq n(n-1)e^{-(\epsilon^2-\epsilon^3)d'/4} \leq n^2 e^{-(\epsilon^2-\epsilon^3)d'/4} < 1,$$

so there exists a linear mapping we want.

This is exactly

$$d' \geq \frac{8\ln n}{\epsilon^2},$$

so we've finally proved JL Lemma. ■

## 11.2 Algorithmic Stability

Though VC theory can give generalization bounds based on hypothesis set used for learning, it ignores the specific algorithm. One may ask if an analysis of the properties of a specific algorithm could lead to finer guarantees. In this section, we will introduce *algorithmic stability* to derive *algorithm-dependent* learning guarantees.

**Definition 11.5 (Uniform stability)** *Let $\mathcal{A}$ be a learning algorithm, $S$ be a training dataset $(z_1, \cdots, z_n)$. Let $S^i = (z_1, \cdots, z_{i-1}, z_i', z_{i+1}, \cdots, z_n)$ denote a* neighbouring dataset *that differ from $S$ by a single point. Let $\mathcal{A}(S)$ denote a model learned from $S$ by $\mathcal{A}$. Let $\ell(\cdot, \cdot)$ be a loss function.*

*The learning algorithm $\mathcal{A}$ is said to have uniform stability $\beta$ with respect to loss $\ell(\cdot, \cdot)$, if $\forall S, \forall i, \forall z_i', \forall z$*

$$|\ell(\mathcal{A}(S), z) - \ell(\mathcal{A}(S^i), z)| \leq \beta$$

Generally, the coefficient $\beta$ depends on the sample size $n$. A learning algorithm is said to be stable if $\beta(n) = O(1/\sqrt{n})$ and to be *very* stable if $\beta(n) = O(1/n)$.

**Definition 11.6 (Risk and Empirical risk)** *Define the risk (similar to test error or generalization error) of an algorithm $\mathcal{A}$ on training dataset $S$ sampled from data distribution $\mathcal{D}$*

$$R(\mathcal{A}(S)) := \mathop{\mathbb{E}}_{z \sim \mathcal{D}} [\ell(\mathcal{A}(S), z)],$$

*and its empirical risk (similar to training error)*

$$R_{\text{emp}}(\mathcal{A}(S)) := \frac{1}{n} \sum_{z_i \in S} \ell(\mathcal{A}(S), z_i)$$

*where $n = |S|$.*

**Lemma 11.7** *If algorithm $\mathcal{A}$ has a uniform stability $\beta$ and is symmetric on $(z_1, \cdots, z_n)$,* i.e. *for any permutation $\sigma$, $\mathcal{A}(z_1, \cdots, z_n) = \mathcal{A}(z_{\sigma(1)}, \cdots, z_{\sigma(n)})$, then*

$$\mathbb{E}_S [R(\mathcal{A}(S)) - R_{\text{emp}}(\mathcal{A}(S))] \leq \beta.$$

**Proof:** For generalization risk,

$$\mathbb{E}_S[R(\mathcal{A}(S))] = \mathbb{E}_S \mathbb{E}_z[\ell(\mathcal{A}(S), z)] = \mathbb{E}_{z_1, \cdots, z_n, z}[\ell(\mathcal{A}(S), z)] = \mathbb{E}_{S, z_1}[\ell(\mathcal{A}(S'), z_1)]$$

where $S' = (z, z_2, \cdots, z_n)$

For empirical risk,

$$\mathbb{E}_S[R_{\text{emp}}(\mathcal{A}(S))] = \mathbb{E}_S \left[ \frac{1}{n} \sum_{i=1}^{n} \ell(A(S), z_i) \right] = \mathbb{E}_{S, z_1}[\ell(\mathcal{A}(S), z_1)]$$

the equality holds according to the symmetry of $\mathcal{A}$.

Thus

$$\mathbb{E}_S [R(\mathcal{A}(S)) - R_{\text{emp}}(\mathcal{A}(S))] = \mathbb{E}_{S, z_1}[\ell(\mathcal{A}(S), z_1) - \ell(\mathcal{A}(S'), z_1)] \leq \beta$$

■

### 11.2.1 Algorithmic Stability of SGD and GD

Deep neural networks have already shown excellent behaviours for generalization. However, if we try to explain it with VC theory, we will find that the VC dimension of deep neural networks is $O(E)$, where $E$ is the number of parameters, and is usually far greater than the number of training data. This will lead to meaningless bounds.

A possible way is to analyze the algorithmic stability of training strategy or optimization algorithm such as GD and SGD.

Intuitively, SGD is a stable algorithm since the randomness of SGD somehow "smooths" the impact of changing a single data. However, we cannot make the same assertion on GD, as it always uses all data to optimize the model, and there exists the possibility that a single data makes a great difference.

In [HRS16], the authors proved bounds on the uniform stability of SGD. In [CP18], the authors proved similar uniform stability of GD in the convex setting. They also constructed an explicit example in a non-convex setting, where GD is not uniformly stable while SGD is.

## 11.3 Proof of Von Neumann's Minmax Theorem through online learning

**Theorem 11.8 (Von Neumann's Minmax Theorem)**

$$\min_p \max_q p^\top Mq = \max_q \min_p p^\top Mq$$

**Proof:**

If two players follow pure strategy, we have (refer to 7.1 in Lecture 7)

$$\min_i \max_j M_{ij} = \max_j \min_i M_{ij}$$

In case of two players adopting mixed strategy, it is also intuitive that playing second is better, because playing second means having more information without any cost. So we have

$$\min_p \max_q p^\top Mq \geq \max_q \min_p p^\top Mq \tag{11.2}$$

We consider row player as an online learner, with each row acting like an expert, and column player as adversary in online learning. So we can obtain an online learning algorithm as Algorithm 1.

**Algorithm 1** Randomized Weighted Majority Vote for Repeated Games

Init: $p_1 = (\frac{1}{n}, ..., \frac{1}{n})$ $M = (m_{ij})_{n\times n}$ $m_{ij} \in [0, 1]$
Param: $\beta \in (0, 1)$
**for** $t = 1, 2, ..., T$ **do**
1) Row player chooses $p_t$
2) Column player (after choosing $p_t$) chooses $q_t$
3) Row player observes $\mathbf{1}_i^\top Mq_t$
4) Update weight $p_{t+1}(i) = p_t(i)\beta^{\mathbf{1}_i^\top Mq_t}/z$
**end for**

So we have

$$\sum_{t=1}^{T} p_t^\top M q_t \leq \frac{\ln \frac{1}{\beta}}{1-\beta} \min_p \sum_{t=1}^{T} p^\top M q_t + \frac{\ln n}{1-\beta}$$

By setting $\beta = \frac{1}{1+\sqrt{\frac{2\ln n}{T}}}$, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} p_t^\top M q_t &\leq \min_p \frac{1}{T}\sum_{t=1}^{T} p^\top M q_t + O(\frac{\log n}{\sqrt{T}}) \\ &\leq \max_q \min_p p^\top M q\end{aligned}$$

Then

$$\begin{aligned}\min_p \max_q p^\top M q &\leq \frac{1}{T}\sum_{t=1}^{T} p_t^\top M q_t \\ &\leq \max_q \min_p p^\top M q\end{aligned}$$

Combined with (11.2), the proof is completed.

■

# References

[HRS16] M. Hardt, B. Recht and Y. Singer, Train faster, generalize better: Stability of stochastic gradient descent, *International conference on machine learning* (2016), pp. 1225-1234.

[CP18] Z. Charles and D. Papailiopoulos, Stability and generalization of learning algorithms that converge to global optima, *International conference on machine learning* (2018), pp. 745-754.