CS 2243 Fall 2024 Scribes: <names>
10/07 Based on notes by: Shi Feng

Lecture 10: Filtered PCA

Overview: This lecture focuses on the problem of learning an unknown ReLU network with respect to Gaussian inputs. We will make several attempts to partially recover the ReLU network in this lecture and finally give an algorithm Filtered PCA with $\mathrm{poly}(d)$ time/samples complexity.

1 Warm-Up

We introduce the feedfoward ReLU networks at first.

Definition 1 (ReLU Networks [CKM22]). *Let $\mathcal{C}_S$ denote the concept class of (feedforward) ReLU networks over $\mathbb{R}^d$ of size S. Specifically, $F \in \mathcal{C}_S$ if there exist weight matrices $\mathrm{W}_1 \in \mathbb{R}^{k_1 \times d}, \mathrm{W}_2 \in \mathbb{R}^{k_2 \times k_1}, \ldots, \mathrm{W}_L \in \mathbb{R}^{1 \times k_{L-1}}$ for which*

$$F(x) \triangleq \mathrm{W}_L \phi\left(\mathrm{W}_{L-1} \phi\left(\cdots \phi(\mathrm{W}_1 x) \cdots\right)\right), \tag{1}$$

where $\phi(z) \triangleq \max(z, 0)$ is the ReLU activation applied entrywise, and $k_1 + \cdots + k_{L-1} = S$. In this case we say that F is computed by a ReLU network with depth L. We make a necessary assumption of Lipschitzness on F such that $|F(x) - F(x')| \le \Lambda \|x - x'\|_2$. Furthermore, it is easy to verify that F is positively homogeneous, i.e., $F(\lambda x) = \lambda F(x), \forall \lambda \ge 0, x \in \mathbb{R}^d$.

Our task is to find an algorithm for learning the network over Gaussian inputs. Λ is required in this task even for one-hidden-layer MLPs. One can refer to [CKM22] for this negative result. Also, in [CKM22], they give a theorem to show the time/sample complexity of Filtered PCA algorithm we will discuss today as below.

Theorem 1 (Theorem 1.2 in [CKM22]). *There is a proper algorithm ("Filtered PCA") for learning MLPs of any depth over Gaussian inputs to error ε in time/samples $\mathrm{poly}(d) \cdot \exp(\mathrm{poly}(S, \Lambda, \log B, 1/\varepsilon))$ where $B = \prod_{i=1}^{L-1} \|\mathrm{W}_i\|_{op}$.*

In this theorem, the complexity is $\mathrm{poly}(d) \cdot \exp(\mathrm{poly}(S))$, which is "fixed parameter tractable" (FPT) compared to $d^{\mathrm{poly}(S)}$ since in $\mathrm{poly}(d) \cdot \exp(\mathrm{poly}(S))$, the factor $\mathrm{poly}(d)$ does not depend on S. Moreover, this task is provably not achievable by correlational statistical query algorithms including tensor methods, kernel methods, noisy gradient descent on square loss.

Before our attempts, we introduce an observation at first. Let $\mathrm{w}_1, \mathrm{w}_2, \cdots, \mathrm{w}_k$ be the weight vectors in the first layer, i.e. the rows of W_1. We call $U = \mathsf{span}(\mathrm{w}_1, \cdots, \mathrm{w}_k)$ the "relevant subspace" because by definition $F(x)$ should only depend on $\Pi_U(x)$. Here, we let $\Pi_U : \mathbb{R}^d \to U$ denote orthogonal projector to U. If one could recover U, then FPT can be got by projecting all data onto U and "brute force" the network in time $(1/\varepsilon)^{\mathrm{poly}(k)}$ without any dependence on d. Hence, our task becomes recovering the space U. CSQ lower bound shows that in the worst case, CSQ algorithms cannot even recover a single vector from the span U of the input weight vectors. However, we will see from the lecture that a "slightly non-CSQ" algorithm can find vectors in U.

1.1 Attempt 1

In the first attempt, we create a matrix $M \triangleq \mathbb{E}[y \cdot S_2(x)]$ where $S_2(x) = xx^\top - \mathrm{I}$. If $F(x) = \sum_{i=1}^k \lambda_i \phi(\langle \mathrm{w}_i, x \rangle)$, which is a MLP with only one hidden layer, we can get $M = \sum_{i=1}^k \lambda_i \mathrm{w}_i \mathrm{w}_i^\top$. If $M \neq 0$, then the top-k singular subpace of M would be U. However, CSQ lower bound implies that $\exists \lambda_i, \mathrm{w}_i, i \in [k]$ such that $M = 0$ and more generally, $\exists \lambda_i, \mathrm{w}_i, i \in [k]$ such that $\sum_{i=1}^k \lambda_i w_i^{\otimes l} = 0$ for $l \in [\lfloor k/2 \rfloor]$. Intuitively, this is because λ_i's can be positive and negative, so the matrix M and even the tensor $\sum_{i=1}^k \lambda_i w_i^{\otimes l}$ can be 0 and provide no information at all.

1.2 Attempt 2

In this attempt, we extend matrix M to M_h such that $M_h \triangleq \mathbb{E}[h(y) \cdot S_2(x)]$ where $h : \mathbb{R} \to \mathbb{R}$ is a "filter" function. We prove a lemma at first.

Lemma 1. *The kernel of M_h, $\ker(M_h)$, contains the orthogonal component $U^\perp$ of U in $\mathbb{R}^d$.*

Proof. In order to prove this, we need to prove that for $\mathrm{w} \in U^\perp$, $M_h\mathrm{w} = 0$. We only need to prove $\mathrm{w}^\top M_h \mathrm{w} = 0$ and $\mathrm{w}'^\top M_h \mathrm{w} = 0$ for any $\mathrm{w}' \perp \mathrm{w}$. Suppose $\|\mathrm{w}\|_2 = \|\mathrm{w}'\|_2$ without loss of generality.

For the first part, we have $\mathrm{w}^\top M_h \mathrm{w} = \mathbb{E}[h(y)(\langle \mathrm{w}, x\rangle^2 - 1)]$. We know that if two vectors $\mathrm{v}, \mathrm{v}' \in \mathbb{R}^d$ and $\mathrm{v} \perp \mathrm{v}'$, $\Pi_{\mathrm{v}}(g)$ and $\Pi_{\mathrm{v}'}(g)$ are independent random variables for any $g \sim \mathcal{N}(0, \mathrm{I})$. Since y depends on x only through $\Pi_U(x)$, whereas $\langle \mathrm{w}, x\rangle^2$ depends on x only through projection to $\mathrm{w} \in U^\perp$, so $h(y)$ and $\langle \mathrm{w}, x\rangle^2$ are independent. Hence, we can deduce that

$$\begin{aligned}
\mathrm{w}^\top M_h \mathrm{w} &= \mathbb{E}[h(y)(\langle \mathrm{w}, x\rangle^2 - 1)] \\
&= \mathbb{E}[h(y)]\mathbb{E}[(\langle \mathrm{w}, x\rangle^2 - 1)] \\
&= \mathbb{E}[h(y)]\mathbb{E}_{g\sim\mathcal{N}(0,1)}[g^2 - 1)] \\
&= 0.
\end{aligned}$$

For the second part, we have $\mathrm{w}'^\top M_h \mathrm{w} = \mathbb{E}[h(y)(\langle \mathrm{w}', x\rangle\langle \mathrm{w}, x\rangle - \langle \mathrm{w}, \mathrm{w}'\rangle)] = \mathbb{E}[h(y)(\langle \mathrm{w}', x\rangle\langle \mathrm{w}, x\rangle)]$ because $\mathrm{w} \perp \mathrm{w}'$. Similarly, because $h(y)\langle \mathrm{w}', x\rangle$ depends on x only through $\Pi_U(x)$, whereas $\langle \mathrm{w}, x\rangle$ depends on x only through projection to $\mathrm{w} \in U^\perp$, so $h(y)\langle \mathrm{w}', x\rangle$ and $\langle \mathrm{w}, x\rangle$ are independent. We can deduce that

$$\begin{aligned}
\mathrm{w}'^\top M_h \mathrm{w} &= \mathbb{E}[h(y)(\langle \mathrm{w}', x\rangle\langle \mathrm{w}, x\rangle)] \\
&= \mathbb{E}[h(y)\langle \mathrm{w}', x\rangle]\mathbb{E}[\langle \mathrm{w}, x\rangle] \\
&= 0.
\end{aligned}$$

Until now, our lemma is proved. □

Using this lemma, we can further prove a corollary.

Corollary 1. *If $M_h \neq 0$, then the top singular vector of M_h lies in U.*

Proof. The top singular vector of M_h should be orthogonal to $\ker(M_h)$ when $M_h \neq 0$. Since $U^\perp \subseteq \ker(M_h)$, we know that the top singular vector of M_h should be orthogonal to $U^\perp$ so in U. □

Therefore, if we can prove $M_h \neq 0$ for some h, we can use this method to find one vector in U. We take $h(y) = y^2$ then we can prove the following lemma.

Lemma 2. *When $h(y) = y^2$, the matrix M_h satisfies $\mathrm{Tr}(M_h) \neq 0$.*

Proof. Since $h(y) = y^2$, we can deduce that

$$\begin{aligned}
\mathrm{Tr}(M_h) &= \mathbb{E}[h(y) \cdot (\|x\|_2^2 - d)] \\
&= \mathbb{E}[y^2(\|x\|_2^2 - d)] \\
&= \mathbb{E}[F(x)^2 \cdot (\|x\|_2^2 - d)].
\end{aligned}$$

We "factorized" $x \sim \mathcal{N}(0, \mathrm{I})$ into two random variables $z \sim \mathcal{S}^{d-1}$ and $r \sim$ distribution of norm of Gaussian vector sampled from $\mathcal{N}(0, \mathrm{I})$ (chi-squared distribution $\mathcal{X}_d^2$). z represents the direction and r represents the magnitude and they are independent by definition. x is given by $\sqrt{r}z$. Hence, we have

$$\begin{aligned}
\mathrm{Tr}(M_h) &= \mathbb{E}[F(x)^2 \cdot (\|x\|_2^2 - d)] \\
&= \mathbb{E}_r[(r^2 - d) \cdot \mathbb{E}_z[F(r \cdot z)^2]] \\
&= \mathbb{E}_r[r^2(r^2 - d) \cdot \mathbb{E}_z[F(z)^2]] && \text{(by positively homogeneous of } F\text{)} \\
&= \mathbb{E}_r[r^2(r^2 - d)]\mathbb{E}_z[F(z)^2].
\end{aligned}$$

We know that $\mathbb{E}_z[F(z)^2] > 0$ because $F \neq 0$. Otherwise, the problem has a trivial solution. For $\mathbb{E}_r[r^2(r^2-d)]$, we have

$$\begin{aligned}\mathbb{E}_r[r^2(r^2-d)] &= \mathbb{E}_r[r^4] - \mathbb{E}_r[r^2]d \\ &= \mathbb{E}_{g_i\sim\mathcal{N}(0,1),\forall i\in[d]}[(g_1^2+g_2^2+\cdots+g_d^2)^2] - d^2 \\ &= \sum_{i\in[d]} \mathbb{E}_{g_i\sim\mathcal{N}(0,1)}[g_i^4] + \sum_{i\neq j} \mathbb{E}_{g_i,g_j\sim\mathcal{N}(0,1)}[g_i^2 g_j^2] - d^2 \\ &= 3d + d(d-1) - d^2 = 2d \neq 0.\end{aligned}$$

This completes our proof. □

Using $h(z) = z^2$ in attempt 2 finds one vector in U and works for the warmup goal, but it is still unclear how to extend it to learn remaining directions in U.

2 Filtered PCA

In this section, we first give a function h that is easier to extend to the case when we want to learn all directions in U. Intuitively, we only need to design a h such that $M_h = \mathbb{E}[h(y)\cdot S_2(x)]$ satisfies that $\langle \Pi_U, M_h\rangle = \mathbb{E}[h(y)\cdot(\|\Pi_U x\|_2^2 - k)]$ is no exponentially small. Intuitively, if this is satisfied, M_h can provide enough information for us to recover all directions in U.

2.1 A Different Filter Function

In order to satisfy the condition, we want h to satisfy another 3 conditions, then we prove that under these 3 conditions, $\langle \Pi_U, M_h\rangle$ is not negligible. The conditions are:

1. $h(z) \geq 0$ for all $z \in \mathbb{R}$.
2. $h(F(x))$ is not identically zero.
3. $h(F(x)) \neq 0$ if and only if $\|\Pi_U(x)\|_2^2 \geq 2k$.

With these 3 conditions, we have

$$\begin{aligned}\langle M_h, \Pi_U\rangle &= \mathbb{E}\left[h(F(x))\cdot(\|\Pi_U x\|_2^2 - k)\cdot\mathbb{I}[\|\Pi_U(x)\|_2^2 \geq 2k]\right] \\ &\quad + \mathbb{E}\left[h(F(x))\cdot(\|\Pi_U x\|_2^2 - k)\cdot\mathbb{I}[\|\Pi_U(x)\|_2^2 < 2k]\right] \\ &= \mathbb{E}\left[h(F(x))\cdot(\|\Pi_U x\|_2^2 - k)\cdot\mathbb{I}[\|\Pi_U(x)\|_2^2 \geq 2k]\right] \\ &\geq \mathbb{E}\left[h(F(x))\cdot k\cdot\mathbb{I}[\|\Pi_U(x)\|_2^2 \geq 2k]\right] \geq 0.\end{aligned}$$

This is likely positive. If we can prove $\langle M_h, \Pi_U\rangle > 0$, according to Corollary 1, the top eigenvector of M_h would lie in the relevant subspace U.

Now we take $h(z) = \mathbb{I}[|z| \geq \tau]$ where $\tau = \Lambda\sqrt{2k}$. Then we have $\langle M_h, \Pi_U\rangle \geq k\cdot\mathbb{E}[h(F(x))\cdot\mathbb{I}[\|\Pi_U(x)\|_2^2 \geq 2k]] = k\cdot\mathbb{E}[\mathbb{I}[|y|\geq\tau]] = k\cdot\mathbb{P}[|F(x)|\geq\tau]$. We only need to prove $\mathbb{P}[|F(x)|\geq\tau] > 0$ to prove $\langle M_h, \Pi_U\rangle > 0$. Moreover, we want to verify the 3 conditions. The first condition is satisfied trivially. The second condition is closely related to lower bound $\mathbb{P}[|F(x)|\geq\tau]$ and we will discuss later. We verify the third condition as below.

Lemma 3 (Condition 3)**.** *When $h(z) = \mathbb{I}[|z|\geq\tau]$, we have $h(F(x)) \neq 0$ if and only if $\|\Pi_U(x)\|_2^2 \geq 2k$.*

Proof. If $\|\Pi_U(x)\|_2^2 < 2k$, we have

$$\begin{aligned}|F(x)| &= |F(\Pi_U(x))| \\ &\leq |F(0)| + \Lambda\|\Pi_U x\|_2 && (\Lambda\text{-lipschitzness of } F) \\ &< \Lambda\sqrt{2k},\end{aligned}$$

so $h(F(x)) = 0$ and the lemma is proved. □

Finally, we focus on the second condition and prove that with high probability, $F(x)| \geq \tau$, which also implies that $\langle M_h, \Pi_U \rangle > 0$. Formally, we have the following lemma.

Lemma 4 (Condition 2)**.** *Suppose $\mathbb{E}_{x\sim\mathcal{N}(0,\mathrm{I})}[F(x)^2] \geq \sigma^2$. Then we can prove that when $h(z) = \mathbb{I}[|z| \geq \tau]$, we have*

$$\mathbb{P}[|F(x)| \geq s] \geq \mathsf{erfc}\left(\frac{\sqrt{2k}s}{\sigma}\right)\frac{\sigma^2}{2k} \geq \Omega(\exp(-3ks^2/\sigma^2))\frac{s\sigma}{\sqrt{k}\Lambda^2}, \tag{2}$$

for any $s > 0$.

Proof. $F : \mathbb{R}^k \to \mathbb{R}$ is a continuous, piecewise-linear function which is Λ-Lipschitz and satisfies $\mathbb{E}_{x\sim\mathcal{N}(0,\mathrm{I})}[F(x)^2] \geq \sigma^2$. Let $S_i \subseteq \mathbb{R}^k$ be a linear piece (polyhedral cone) of F. Suppose $F(x) = \langle u_i, x\rangle$ for $\forall x \in S_i$. Without loss of generality, we assume that $\|u_i\|_2 \leq \Lambda$ (see lemma 4.5 in [CKM22]). We define $\sigma_i^2 \triangleq \mathbb{E}_{x\sim\mathcal{N}(0,\mathrm{I})}[\langle u_i, x\rangle^2 | x \in S_i]$.

Note that if we choose linear piece S_i with probability $\mathbb{P}[x \in S_i]$, then $\mathbb{E}[F(x)^2] = \mathbb{E}_i[\sigma_i^2] \geq \sigma^2$. Because S_i is a polyhedral cone, similarly to the proof of Lemma 2, sampling x from $\mathcal{N}(0, \mathrm{I})$ given $x \in S_i$ is equivalent to sampling r from $\mathcal{X}_k^2$ and v from $\mathcal{S}^{k-1}$ given $v \in S_i$ and then outputting $\sqrt{r}v$. Hence, we have

$$\begin{aligned}
\sigma_i^2 &= \mathbb{E}_{r\in\mathcal{X}_k^2, v\in\mathcal{S}^{k-1}}[r\langle u_i, v\rangle^2 | v \in S_i] \\
&= \mathbb{E}_r[r] \cdot \mathbb{E}_v[\langle u_i, v\rangle^2 | v \in S_i] \\
&= k \cdot \mathbb{E}_v[\langle u_i, v\rangle^2 | v \in S_i],
\end{aligned}$$

so we have $\mathbb{E}_v[\langle u_i, v\rangle^2 | v \in S_i] = \frac{\sigma_i^2}{k}$. Before we continue our proof, we prove a claim at first.

Claim 1. *If a random variable Z satisfies that $|Z| \leq M$ almost surely and $\mathbb{E}[Z^2] \geq \sigma^2$, we have $\mathbb{P}[|Z| \geq t] \geq \frac{1}{M^2}(\sigma^2 - t^2)$.*

Proof. We can deduce that

$$\begin{aligned}
\sigma^2 &\leq \mathbb{E}[Z^2] \\
&= \mathbb{E}[Z^2 | |Z| \geq t] \cdot \mathbb{P}[|Z| \geq t] + \mathbb{E}[Z^2 | |Z| < t] \cdot \mathbb{P}[|Z| < t] \\
&\leq M^2 \cdot \mathbb{P}[|Z| \geq t] + t^2,
\end{aligned}$$

which indicates that $\mathbb{P}[|Z| \geq t] \geq \frac{1}{M^2}(\sigma^2 - t^2)$. □

According to the claim, we can further deduce that

$$\begin{aligned}
\mathbb{P}[|\langle u_i, x\rangle| \geq s | v \in S_i] &\geq \mathbb{P}\left[r \geq \frac{2ks^2}{\sigma_i^2}\right] \cdot \mathbb{P}\left[|\langle u_i, v\rangle| \geq \frac{\sigma_i}{\sqrt{2k}} \middle| v \in S_i\right] \\
&\geq \mathbb{P}\left[r \geq \frac{2ks^2}{\sigma_i^2}\right] \cdot \left(\frac{\sigma_i^2}{k} - \frac{\sigma_i^2}{2k}\right) \\
&\geq \mathbb{P}_{g\in\mathcal{N}(0,\mathrm{I})}[g \geq \sqrt{2k}s/\sigma_i] \cdot \frac{\sigma_i^2}{2k} \\
&= \mathsf{erfc}\left(\frac{\sqrt{2k}s}{\sigma_i}\right) \cdot \frac{\sigma_i^2}{2k}.
\end{aligned}$$

We can verify that $\mathsf{erfc}\left(\frac{\sqrt{2k}s}{\sigma_i}\right)\cdot\frac{\sigma_i^2}{2k}$ is a convex function with respect to σ_i. Hence, we have

$$\begin{aligned}
\mathbb{P}[|F(x)|\geq s] &= \mathbb{E}_i[\mathbb{P}[|\langle u_i,x\rangle|\geq s|v\in S_i]] \\
&\geq \mathsf{erfc}\left(\frac{\sqrt{2k}s}{\mathbb{E}[\sigma_i]}\right)\cdot\frac{\mathbb{E}_i[\sigma_i^2]}{2k} && \text{(Jensen's inequality)}\\
&\geq \mathsf{erfc}\left(\frac{\sqrt{2k}s}{\mathbb{E}[\sigma_i^2]^{1/2}}\right)\cdot\frac{\sigma^2}{2k} \\
&= \mathsf{erfc}\left(\frac{\sqrt{2k}s}{\sigma}\right)\frac{\sigma^2}{2k}.
\end{aligned}$$

The second inequality in this lemma is proved by standard bounds on erfc. □

Now we have proved that $h(z)=\mathbb{I}[|z|\geq\tau]$ satisfies all three conditions and M_h has kernel containing $U^\perp$ and top eigenvalue no less than $\exp(-O(\Lambda^2k^2/\sigma^2))$ (let $s=\tau$ in Lemma 4). If we form an empirical estimate of M_h from enough samples and take its top eigenvector, will be close to the relevant subspace U. However, The full spectrum of M_h need not reveal the full subspace U. The next question is how to extend our method to learn the rest of the relevant subspace.

2.2 Finding Another Relevant Vector

Given $v\in U$, we want to find another vector $v'\in U\backslash\mathsf{span}(v)$. We define a different matrix M' as

$$M'\triangleq\Pi_v^\perp\cdot\mathbb{E}[\mathbb{I}[|y-F(\Pi_v x)|\geq\tau]\cdot(xx^\top-\mathrm{I})]\cdot\Pi_v^\perp. \tag{3}$$

Here we suppose we already know $F(\Pi_v x)$. In fact, if F is an MLP with size no more than S over a known subspace V, we can compute $F(\Pi_v x)$ by constructing an ε-net over possible networks (problem 1 in pset 4).

Similar to Lemma 1, we claim that $\ker(M')$ contains $U^\perp\oplus\mathsf{span}(v)$ as below.

Lemma 5. *The kernel of M', $\ker(M')$ contains subspace $U^\perp\oplus\mathsf{span}(v)$ of $\mathbb{R}^d$.*

Proof. For $z=\alpha v+\beta w$ where $w\in U^\perp$ is unit form and $v\perp w$, we have

$$\begin{aligned}
z^\top M'z &= \beta^2 w^\top\cdot\mathbb{E}[\mathbb{I}[|y-F(\Pi_v x)|\geq\tau]\cdot(xx^\top-\mathrm{I})]\cdot w \\
&= \beta^2\mathbb{E}[\mathbb{I}[|y-F(\Pi_v x)|\geq\tau]\cdot(\langle w,x\rangle^2-1)] \\
&= \beta^2\mathbb{E}[\mathbb{I}[|y-F(\Pi_v x)|\geq\tau]]\mathbb{E}[\langle w,x\rangle^2-1] \\
&= 0.
\end{aligned}$$

Here, the second last equality is because $\mathbb{I}[|y-F(\Pi_v x)|\geq\tau]$ depends on $U\oplus v$ and $\langle w,x\rangle^2-1$ depends on w, which is orthogonal to $U\oplus v$. □

Similar to Corollary 1, we can also prove a corollary using this lemma.

Corollary 2. *If $M'\neq 0$, then the top singular vector of M' lies in $U\backslash\mathsf{span}(v)$.*

Proof. The top singular vector of M' should be orthogonal to $\ker(M')$ when $M'\neq 0$. Since $(U^\perp\oplus\mathsf{span}(v))\subseteq\ker(M')$, we know that the top singular vector of M' should be orthogonal to $U^\perp\oplus\mathsf{span}(v)$ so in $U\backslash\mathsf{span}(v)$. □

Similar to before, we will prove $M'\neq 0$ by lower bounding $\langle M',\Pi_{U\backslash\mathsf{span}(v)}\rangle$. In detial, we have

$$\langle M',\Pi_{U\backslash\mathsf{span}(v)}\rangle=\mathbb{E}[\mathbb{I}[|y-F(\Pi_v x)|\geq\tau]\cdot(\left\|\Pi_{U\backslash\mathsf{span}(v)}(x)\right\|_2^2-(k-1))],$$

and similar to Section 2.1, we need to check the two claims:

1. $\mathbb{P}[|y-F(\Pi_v(x))|\geq\tau]$ is large (corresponding to condition 2 in Section 2.1 proved by Lemma 4).

2. $|y - F(\Pi_v(x))| \geq \tau$ for $x \perp v$ if and only if $\left\|\Pi_{U\backslash \mathsf{span}(v)}(x)\right\|_2^2 \geq 2k-2$ (corresponding to condition 3 in Section 2.1 proved by Lemma 3).

Then we can deduce that $\langle M, \Pi_{U\backslash \mathsf{span}(v)}\rangle \geq (k-1)\cdot \mathbb{P}[|y - F(\Pi_v(x))| \geq \tau] > 0$, which indicates that the top eigenvector of M' lies in $U\backslash \mathsf{span}(v)$.

We still use $h(z) = \mathbb{I}[|z| \geq \tau]$ but change τ from $\Lambda\sqrt{2k}$ to $\Lambda\sqrt{2k-2}$. We prove condition 2 at first.

Lemma 6 (Condition 2)**.** *When $h(z) = \mathbb{I}[|z| \geq \tau]$, we have $h(F(x)) \neq 0$ if and only if $\|\Pi_U(x)\|_2^2 \geq 2k$.*

Proof. Take any $x \perp v$. If $\left\|\Pi_{U\backslash \mathsf{span}(v)}(x)\right\|_2^2 < 2k-2$, then $|F(x) - F(\Pi_v(x))| < \Lambda\sqrt{2k-2}$:

$$\begin{aligned}|F(x) - F(\Pi_v(x))| &= |F(\Pi_U(x)) - F(\Pi_v(x))| \\ &\leq \Lambda \left\|\Pi_{U\backslash \mathsf{span}(v)}(x)\right\|_2 \\ &\leq \Lambda\sqrt{2k-2},\end{aligned}$$

so condition holds for $\tau = \Lambda\sqrt{2k-2}$. □

Then we prove condition 1, $\mathbb{P}[|y - F(\Pi_v(x))| \geq \tau]$ is large. On one hand, if $\mathbb{E}[(y - F(\Pi_v(x)))^2]$ is large, because $F(x) - F(\Pi_v(x))$ is piecewise-linear, we can apply anti-concentration argument from earlier (you can prove it similarly to Lemma 4 if you are interested). On the other hand, if $\mathbb{E}[(y - F(\Pi_v(x)))^2]$ is small, $F(\Pi_v(x))$ already achieves low test loss, so we are done! Formally, we have the following more general lemma, which is similar to Lemma 4. suppose we have found orthogonal vectors $v_1, \cdots, v_m \in U$ so far. Let V denote their span, then we have the following lemma to show that we can find the next $v_{m+1} \in U\backslash V$ if $m < k$.

Lemma 7 (A More General Condition 1, Lemma 5.5 in [CKM22])**.** *If $\mathbb{E}[(y - F(\Pi_v(x)))^2] \geq \sigma^2$ and $\tau = \Lambda\sqrt{2(k-m)}$, then*

$$M' \triangleq \Pi_{V^\perp} \cdot \mathbb{E}[\mathbb{I}[|y - F(\Pi_V(x))| \geq \tau] \cdot S_2(x)] \cdot \Pi_{V^\perp}$$

has kernel containing $U^\perp \oplus V$ and top eigenvalue no less than $\exp(-O(\Lambda^2 k^2/\sigma^2))$.

When $m = 1$, the lemma is condition 1 so we can find another relevant vector in $U\backslash \mathsf{span}(v)$. Furthermore, this lemma indicates that if we form an empirical estimate of M' for each m from enough samples and take its top eigenvector, we recover a new direction in U, so we can find all $v_1, v_2, \cdots, v_k$ such that U can be fully recovered. The proof and more strict description of this lemma can be found in [CKM22].

2.3 The Algorithm

In Section 2.1 and Section 2.2, we have already provided a method to recover relevant space U. In this section, we write the whole algorithm Filtered PCA more clearly in an algorithm environment (Algorithm 1).

3 Robustness of Filtered PCA

Let us look back $F(\Pi_V(x))$. In all analysis in Section 2.1 and Section 2.2, we assume that $F(\Pi_V(x))$ is known. However, we actually do not have access to the function. In Section 2.2, we suggest using ε-net to estimate $F(\Pi_V(x))$. In this section, we discuss the robustness of Filtered PCA over the uncertainty of the estimation $\hat{F}(\Pi_V(x))$ of $F(\Pi_V(x))$. To understand how robust our approach is to this, need to understand stability of thresholds of MLPs. We have the following lemma for this.

Lemma 8 (Lemma 2.2 in [CKM22])**.** *For depth-L MLPs F and F' with the same architecture and whose weight matrices are α-close in operator norm, $\mathbb{P}[|F(x)| > \tau, |F'(x)| \leq \tau] \leq 2^{O(S)} B\alpha/\tau$ where α is the granularity of the epsilon-net.*

Algorithm 1: FILTERED PCA

Input: Sufficiently many data points $x \sim \mathcal{N}(0, \mathrm{I})$ and $F(x)$
Output: Set $\mathcal{L}$ of k orthogonal vectors in relevant space U

$\mathcal{L} \leftarrow \emptyset$
for $0 \le m < k$ **do**
  $V \leftarrow \mathsf{span}(\text{vectors in } \mathcal{L})$
  for $\hat{F} \in \varepsilon$-*net(size*$\le S$ *networks over* V*)* **do**
    If $\hat{F}$ achieves low square loss, return $\hat{F}$.
    Else, compute top eigenvector v_{m+1} of $M \triangleq \Pi_{V^\perp} \cdot \mathbb{E}[\mathbb{I}[|y - \hat{F}(\Pi_V(x))| \ge \tau] \cdot S_2(x)] \cdot \Pi_{V^\perp}$.
    If spectral norm of M sufficiently large, continue to 2c.
  end
  $\mathcal{L} \leftarrow \mathcal{L} \cup \{v_{m+1}\}$
end
Return $\mathcal{L}$

Proof. Here we only give a proof sketch. Firstly, we know that any continuous piecewise-linear function F can be written as a depth-2 max-min formula [Ovc00]. We consider max-min formulas Φ, Φ' for F, F'. Then if MLPs F, F' have same architecture, Φ, Φ' have same clause structure. We consider a sequence of "hybrids" between Φ, Φ' changing Φ from $\Phi = \Phi^(0)$, $\Phi^{(1)}$, ..., to $\Phi^{(N)} = \Phi'$ little by little. Between consecutive hybrids $\Phi^{(i-1)}$ and $\Phi^{(i)}$, the probability that their thresholds disagree is at most $\mathbb{P}[\mathbb{I}[\langle u_i, x\rangle > \tau], \mathbb{I}[\langle w_i, x\rangle \le \tau]]$ where $\langle w_i, x\rangle$'s belong to Φ and $\langle u_i, x\rangle$'s belong to Φ'. □

This lemma indicates that when α is very small, the difference between F and F' with respect to τ will be bounded by $2^{O(S)} B\alpha/\tau$. Thus M we used in Algorithm 1 should be very close to M' we use in analysis in Section 2.2. Intuitively, the subspace distance between $\mathsf{span}(v_1, \cdots, v_k)$ output by Filtered PCA and U should be also very small. One who is interested can refer to [CKM22] for more formal claims.

Moreover, we want to emphasis that the linear dependence on α is crucial. If we instead had a bound scaling with $\alpha^{0.99}$, then if we iterate this $\dim(U) = k$ times, would yield sample complexity doubly exponential in k.

4 Takeaways

Known that CSQ algorithms (based only on statistics of the form $\mathbb{E}[y \cdot g(x)]$) fail even to learn general one-hidden-layer MLPs over Gaussian inputs.

We circumvented this by using non-CSQ statistics of the form $\mathbb{E}[h(y - f(x)) \cdot g(x)]$, giving a $\mathsf{poly}(d) \cdot \exp(\mathsf{poly}(\text{problem parameters})/\varepsilon)$-time algorithm for learning general MLPs over Gaussian inputs.

References

[CKM22] Sitan Chen, Adam R. Klivans, and Raghu Meka. Learning deep relu networks is fixed-parameter tractable. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 696–707, 2022.

[Ovc00] Sergei Ovchinnikov. Max-min representation of piecewise linear functions. *arXiv preprint math/0009026*, 2000.