10/18/23

Lecture 14: Filtered PCA

Filtered PCA (warmup):

$$f(x) = W_L \, \mathrm{ReLU}(W_{L-1} \cdots \cdots \mathrm{ReLU}(W_2 \, \mathrm{ReLU}(W_1 x)) \cdots)$$

Let $v_1, \ldots, v_k$ be rows of W_1.

Warmup goal: find vector in $\mathrm{span}(v_1, \ldots, v_k) := V$.

(Let $\Pi_V : \mathbb{R}^d \to V$ denote orthogonal projector to V.)

Attempt 1:

$$\mathbb{E}\Big[y \cdot \underbrace{(xx^T - Id)}_{= S_2(x)} \Big] = 2 \sum_i \lambda_i v_i v_i^T,$$

so if this were nonzero, then top-k singular subspace would be V

What to do when $\mathbb{E}[y \cdot (xx^T - Id)] = 0$?

Attempt 2:

Consider

$$\mathbb{E}[h(y) \cdot (xx^T - Id)] := M_h$$

for "filter" function $h: \mathbb{R} \to \mathbb{R}$.

Claim: If $w \perp V$, then $w^T M_h w = 0$

Pf: Suppose WLOG $\|w\|_2 = 1$

$$w^T \mathbb{E}\left[h(y) \, (xx^T - Id) \right] w$$

$$= \mathbb{E}\left[h(y) \left(\langle w, x \rangle^2 - 1 \right) \right]$$

y depends on x only through $\Pi_V x$, whereas $\langle w, x \rangle^2$ depends on x only through projection to $w \perp V$, so $h(y)$ and $\langle w, x \rangle^2$ are independent!

$$= \mathbb{E}\left[h(y) \right] \cdot \underbrace{\mathbb{E}\left[\langle w, x \rangle^2 - 1 \right]}_{\mathbb{E}_{g \sim N(0,1)} \left[g^2 - 1 \right] = 0}$$

$$= 0.$$

□

Cor: If $M_h \neq 0$, then top singular vector lies in V.

Pf: Top singular vec is orthogonal to $\ker(M_h)$ because $M_h \neq 0$. Also $V^\perp \subseteq \ker(M_h)$. So top singular vec orthogonal to $V^\perp$ and thus lies in V.

□

So suffices to find $h: \mathbb{R} \to \mathbb{R}$ s.t.

$$M_h \neq 0 \,!$$

$\boxed{\text{Sufficient condition. } \operatorname{Tr}(M_h) \neq 0}$

$$\operatorname{Tr}(M_h) = \mathbb{E}\left[h(y) \cdot \left(\|x\|^2 - d\right)\right]$$

$\boxed{\text{Take } h(y) = y^2}$

$$= \mathbb{E}\left[y^2 \cdot \left(\|x\|^2 - d\right)\right]$$

$$= \mathbb{E}\left[f(x)^2 \cdot \left(\|x\|^2 - d\right)\right]$$

$x \sim N(0, I_d)$ can be "factorized" into

$z \sim \mathbb{S}^{d-1}$ and $r \sim$ [norm of Gaussian vector]

(note, x and z independent)

$$= \mathbb{E}_r\left[\left(r^2 - d\right) \cdot \mathbb{E}_z\left[f(r \cdot z)^2\right]\right]$$

b/c $f(r \cdot z) = r \cdot f(z)$ by homogeneity of ReLU networks

$$= \mathbb{E}_r\left[r^2\left(r^2 - d\right) \mathbb{E}_z\left[f(z)^2\right]\right]$$

$$= \underbrace{\mathbb{E}_r\left[r^2\left(r^2 - d\right)\right]} \cdot \underbrace{\mathbb{E}_z\left[f(z)^2\right]}_{>0 \text{ because } f \neq 0}$$

$$= \mathbb{E}[r^4] - \underbrace{\mathbb{E}[r^2]}_{d^2} \cdot d$$

$$= \mathbb{E}_{g \sim N(0, I_d)}\left[(g_1^2 + \dots + g_d^2)^2\right] - d^2$$

$$= \sum_i \mathbb{E}(g_i^4) + \sum_{i \neq j} \mathbb{E}(g_i^2 g_j^2) - d^2$$

$$= 3d + d(d-1) - d^2 = 2d$$

Attempt 2 works for warmup goal, but unclear how to extend it to learn remaining directions in V.

<u>Attempt 3</u>: instead of $h(y) = y^2$, take

$$\boxed{h(y) = \mathbb{1}[|y| > \tau]}$$

for threshold $\tau > 0$.

In lecture slides, showed that it suffices to prove $\Pr[|f(x)| > \tau]$ not too small (anti-concentration).

<u>Lemma</u>: For $f : \mathbb{R}^k \to \mathbb{R}$ a continuous, piecewise-linear function which is Λ-Lipschitz and satisfies $\mathbb{E}_{x \sim N(0, I_d)}[f(x)^2] \geq \sigma^2$,

$$\Pr[|f| > s] \geq \Omega\left(\exp(-3ks^2/\sigma^2)\right) \cdot \frac{s\sigma}{\sqrt{k}\,\Lambda^2}.$$

(polyhedral cone)

Pf: Let $S_i \subseteq \mathbb{R}^k$ be a linear piece of f; Suppose $f(x) = \langle u_i, x\rangle \ \forall x \in S_i$. Can assume WLOG $\|u_i\| \le \Lambda$ (see Lemma 4.5 in [Chen-Klivans-Meka '20]). Define

$$\sigma_i^2 = \mathbb{E}_{x \sim N(0, Id)}\left[\langle u_i, x\rangle^2 \mid x \in S_i\right]$$

Note if linear piece chosen w/ prob $\Pr[x \in S_i]$, then

$$\mathbb{E}\left(f(x)^2\right) = \mathbb{E}_i\left[\sigma_i^2\right] \ge \sigma^2.$$

Because S_i is polyhedral cone,

chi-squared dist. w/ k degrees of freedom

sampling $x \sim N(0, Id) \mid x \in S_i \iff$ - sampling $r \sim \chi_k^2$,
- sampling $v \sim \mathbb{S}^{k-1} \mid v \in S_i$,
- outputting $\sqrt{r} \cdot v$

$$\text{so } \sigma_i^2 = \mathbb{E}_{r,v}\left[r\langle u_i, v\rangle^2 \mid v \in S_i\right]$$

$$= \mathbb{E}_r[r] \cdot \mathbb{E}_v\left[\langle u_i, v\rangle^2 \mid v \in S_i\right]$$

$$= k \cdot \mathbb{E}_v\left[\langle u_i, v\rangle^2 \mid v \in S_i\right],$$

$$\text{so } \mathbb{E}_v\left[\langle u_i, v\rangle^2 \mid v \in S_i\right] = \frac{\sigma_i^2}{k}$$

Claim: If random variable Z satisfies

1) $|Z| \leq M$ almost surely
2) $\mathbb{E}[Z^2] \geq \sigma^2$,

$\Pr[|Z| > t] \geq \frac{1}{M^2}(\sigma^2 - t^2)$

Pf: $\sigma^2 \leq \mathbb{E}[Z^2] = \mathbb{E}[Z^2 \mid |Z| \geq t] \cdot \Pr(|Z| \geq t]$

$+ \mathbb{E}[Z^2 \mid |Z| < t] \cdot \Pr[|Z| < t]$

$\leq M^2 \cdot \Pr[|Z| \geq t] + t^2$

$\Rightarrow \Pr[|Z| \geq t] \geq \frac{\sigma^2 - t^2}{M^2}$ as claimed. $\square$

So $\Pr[|\langle u_i, x\rangle| \geq s \mid v \in S_i]$ (✡)

$\geq \Pr\left[r \geq \frac{2k s^2}{\sigma_i^2}\right] \cdot \Pr\left[|\langle u_i, v\rangle| \geq \frac{\sigma_i}{\sqrt{2k}} \mid v \in S_i\right]$

$\geq \Pr\left[r \geq \frac{2k s^2}{\sigma_i^2}\right] \cdot \left(\frac{\sigma_i^2}{k} - \frac{\sigma_i^2}{2k}\right)$

$\geq \Pr_{g \sim N(0,1)}\left[g \geq \frac{\sqrt{2k}\, s}{\sigma_i}\right] \cdot \frac{\sigma_i^2}{2k}$

$= \mathrm{erfc}\left(\frac{\sqrt{2k}\, s}{\sigma_i}\right) \cdot \frac{\sigma_i^2}{2k}$

can check this is convex

$$\Pr\left[\,|f(x)| \geq s\,\right] = \mathop{\mathbb{E}}_{i}\left[(*)\right]$$

(Jensen's)

$$\geq \operatorname{erfc}\left(\frac{\sqrt{2k}\, s}{\mathop{\mathbb{E}}_{i}[\sigma_i]}\right) \cdot \frac{\overbrace{\mathop{\mathbb{E}}_{i}[\sigma_i^2]}^{=\sigma^2}}{2k}$$

$$\geq \operatorname{erfc}\left(\frac{\sqrt{2k}\; s}{\underbrace{\mathop{\mathbb{E}}_{i}[\sigma_i^2]^{1/2}}_{=\sigma}}\right) \frac{\sigma^2}{2k}.$$

Lemma follows by standard bounds on $\operatorname{erfc}(\cdot)$. $\square$