Compression

$$H=\{\mathbb{R}^d \ni x \mapsto y = W_N\sigma(\cdots W_2\sigma(W_1x)\cdots) \in \mathbb{R}^d : W_n \in \mathbb{R}^{d \times d}, n \in [N]\}$$

Assume σ is γ-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:

$$H_r = \{x \to y = U_N V_N^T \sigma(U_{N-1}V_{N-1}^T \cdots \sigma(U_1V_1^T x)\cdots) : U_n, V_n \in \mathbb{R}^{d \cdot r}\ n \in [N]\}$$

Assume that each of the $2Ndr$ $\underbrace{}_{\text{Parameters}}$ 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 $\forall y, \hat{y}, y'$ $|\ell(y,\hat{y}) - \ell(y,y')| \le \beta \|\hat{y} - y'\|$

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 represneting H_r using b bits, thus

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

In addition, $\ell(\cdot,\cdot)$ is β-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 $S \sim D^m$:

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

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):

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

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 $\|W_n'\|_{specteral} = \|W_n\|_{specteral}\ \forall n$

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 $|\sigma_1| \geq |\sigma_2| \geq \dots \geq |\sigma_d|$ then the best rank-r approximation to W_n is $W_n' = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, thus we can see clearly that $\|W_n'\|_{specteral} = |\sigma_1| = \|W_n\|_{specteral}$ and thus the bound denote above holds for each $r \in [d]$.

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

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

$\Rightarrow$ In total we'll get $\forall r \in [d]$ w.p $\geq 1-\delta$ over $S \sim D^m$:

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

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 $1 \leq r \leq d$:

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

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

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

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

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

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

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

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

$\square$