Foundations of Deep Learning **2020-2021 Spring Semester**

Assignment 4: Generalization

Due: June 28, 2021

Important Guidelines

1. Work can be done in pairs.
2. Solution should be submitted as a single .zip file containing: (i) a PDF that includes answers to theoretical questions and reports for experimental tasks; and (ii) all source files required to reproduce experimental results. It is highly recommended to typeset the solutions in LaTeX or Word. If you choose to submit scanned handwritten solutions, please make sure they are clearly written and the scan is of high quality.

Part 1: Empirical phenomena

1. **Experiment (35 points):** Demonstrate the four empirical postulates we relied on in class when rationalizing about generalization in deep learning.

   It is recommended to work with a common classification dataset (*e.g.* CIFAR10) and use a standard architecture (*e.g.* InceptionV3 network). In particular, you may take inspiration from [1], but are free to choose the exact setting as you wish. To speed up training it is strongly recommended to run the experiments on a GPU (*e.g.* on Google Colab).

Part 2: Generalization Bounds

Compression

1. **(13 points)** Consider a feed-forward fully connected neural network with input, hidden and output dimensions all equal to d:

   $$\mathcal{H} = \left\{ \mathbb{R}^d \ni x \mapsto y = W_N \sigma(W_{N-1}\sigma(W_{N-2}\sigma(\dots W_2\sigma(W_1 x))\dots) \in \mathbb{R}^d \ : \ W_n \in \mathbb{R}^{d,d}, n \in [N] \right\} .$$

   Assume the point-wise activation $\sigma(\cdot)$ is γ-Lipschitz, and satisfies $\sigma(0) = 0$. Assume also that the input domain consists of vectors with Euclidean norm 1 or less. For $r \in [d]$, let $\mathcal{H}_r$ be the hypotheses space corresponding to the same network as above, when its weight matrices are constrained to have rank r or less, i.e.:

   $$\mathcal{H}_r = \left\{ x \mapsto y = U_N V_N^T \sigma(U_{N-1} V_{N-1}^T \sigma(\cdots \sigma(U1 V_1^T x)) \cdots) \ : \ U_n, V_n \in \mathbb{R}^{d,r}, n \in [N] \right\}$$

   Assume that each of the $2Ndr$ parameters representing $\mathcal{H}_r$ is stored in memory using b bits (e.g. $b = 32$). Given a loss function $\ell : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$ that is ρ-Lipschitz in its second argument, i.e. that meets $|\ell(y,\hat{y}) - \ell(y,\hat{y}')| \leq \rho \cdot \|\hat{y} - \hat{y}'\| \ \forall y, \hat{y}, \hat{y}' \in \mathbb{R}^d$, we would like to derive generalization bounds (bounds on gap between population and empirical losses) for $\mathcal{H}$.

   (a) Fix $r \in [d]$, and derive a generalization bound for $\mathcal{H}$ by compressing it into $\mathcal{H}_r$.

   (b) Derive a generalization bound for $\mathcal{H}$ by simultaneously compressing it into $\mathcal{H}_r$ for all $r \in [d]$.

Radamacher complexity and norms

1. **(13 points)** Let $\mathcal{H} = \{h_\theta : \mathcal{X} \to \mathcal{Y} \ : \ \theta \in \mathbb{R}^p, \|\theta\|_\infty \leq 0.5\}$ be a hypotheses 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 $\mathcal{H}_\Theta := \{h_\theta : \theta \in \Theta\} \subseteq \mathcal{H}$. Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$ and a training set $S = \{(x_i, y_i)\}_{i=1}^m \subseteq \mathcal{X} \times \mathcal{Y}$, the Rademacher complexity of $\mathcal{H}_\Theta$ is defined to be:

   $$R(\ell \circ \mathcal{H}_\Theta \circ S) := \frac{1}{m}\mathbb{E}_{\xi_1,\xi_2,\ldots,\xi_m \overset{i.i.d}{\sim} \begin{cases}+1, & \text{w.p. } 0.5\\ -1, & \text{w.p. } 0.5\end{cases}} \left[\sup\nolimits_{v \in \ell \circ \mathcal{H}_\Theta \circ S} \sum\nolimits_{i=1}^m \xi_i v_i\right]$$

   where:

   $$\ell \circ \mathcal{H}_\Theta \circ S := \{(\ell(y_1, h(x_1)), \ell(y_2, h(x_2)), \ldots, \ell(y_m, h(x_m)) : h \in \mathcal{H}_\Theta\} \subseteq \mathbb{R}^m$$

   Assume that:

   $$\mathbb{E}_S[R(\ell \circ \mathcal{H}_\Theta \circ S)] = Volume(\Theta) := \int_{\theta \in \mathbb{R}^p} \mathbb{1}[\theta \in \Theta] d\theta$$

   Assume also that the implicit regularization of optimization leads to solutions with *high* $\|\cdot\|_\infty$, i.e. to:

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

   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.

PAC-Bayes

1. **(13 points)** Prove the following lemma presented in class.

   Lemma: Consider two multivariate Gaussian distributions over $\mathbb{R}^r$–$\mathcal{N}(\mu_0, \Sigma_0)$ and $\mathcal{N}(\mu_1, \Sigma_1)$, where Σ_0 and Σ_1 are non-singular (positive definite). It holds that:

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

2. **(13 points)** In class we used PAC-Bayes to derive a generalization bound that guarantees small population loss when the learned solution is a flat minimum and has low (Euclidean) norm. Suppose we know that the implicit regularization of optimization indeed tends to find 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 $\{\theta_1, \theta_2, \ldots, \theta_k\}$. Using PAC-Bayes, derive a generalization bound that accounts for the latter implicit regularization.

Part 3: Implicit Regularization

Linear regression

1. **(13 points)** In class we proved the following result for the implicit regularization in underdetermined linear regression.

Proposition: With the notations and setting established in class, suppose we minimize $L_S(w)$ by initializing $w^{(0)} = 0$, and producing iterates $w^{(1)}, w^{(2)}, w^{(3)}, \ldots$ via iterative algorithm in which every update $w^{(t+1)} - w^{(t)}$ lies in $span\{\nabla \ell_{(x_i,y_i)}(w) : i \in [m], w \in \mathbb{R}^d\}$ (this includes as special cases gradient descent, stochastic gradient descent and momentum). Assume convergence to global optimum, i.e. to a solution with zero loss. The, this global optimum is the one with minimum (Euclidean) norm.

Generalize this result by proving that if the zero initialization is replaces by an arbitrary initialization $w^{(0)} = a \in \mathbb{R}^d$, 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 $\leq \|P_\perp a\|$, where $P_\perp : \mathbb{R}^d \to \mathbb{R}^d$ stands for projection onto the orthogonal complement of $span\{x_i\}_{i=1}^m$.

Matrix factorization

1. **(Bonus 10 points)** Consider a matrix completion setting, where we observe all entries of a ground truth matrix $W^* \in \mathbb{R}^{d,d'}$, meaning the training loss is $L_S(W) = \frac{1}{d \cdot d'} \cdot \frac{1}{2} \cdot \|W - W^*\|_{Fro}^2$. Let $W^* = U \Sigma V^T$ be a singular value decomposition of W^*, i.e. $U \in \mathbb{R}^{d,d}$ and $V \in \mathbb{R}^{d',d'}$ are orthogonal, and $\Sigma \in \mathbb{R}^{d,d'}$ is rectangular-diagonal with non-negative entries. Suppose we optimize $L_S(\cdot)$ by overparameterization with a depth N linear neural network and running gradient flow starting from balanced initialization. Assume that at initialization, the end-to-end matrix is given by $W_{1:N}(t) = U \mathcal{E} V^T$, where $\mathcal{E} \in \mathbb{R}^{d,d'}$ is a rectangular-diagonal matrix with diagonal entries all equal to some $\epsilon > 0$, which is much smaller than all singular values of W^*. It can be shown that in this case the analytic singular values decomposition of $W_{1:N}(t)$ is given by $W_{1:N}(t) = US(t)V^T$, i.e. the left and right singular vectors are fixed. For depths $N = 1$ and $N = 2$, derive explicit expressions for $\sigma_1(t), \sigma_2(t), \ldots, \sigma_{min\{d,d'\}}(t)$ – singular values of $W_{1:N}(t)$. Explain how added depth (case $N = 2$) leads singular values to "shoot up" one by one, in line with the interpretation given in class.

References

[1] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations (ICLR)*, 2017.