Foundations of Deep Learning **Fall 2024**

Exercise 8: Generalization II

Lecturer: Aurelien Lucchi

Problem 1 (Vapnik-Chervonenkis Theory):

To quantify the effective 'size' or 'richness' of an infinite class of functions, we use the concept of *shattering coefficient* (aka *growth function*). The **shattering coefficient** of a binary function class $\mathcal{F}$ is defined as the maximum number of ways into which n points can be classified by the function class $\mathcal{F}$

$$\mathcal{S}_{\mathcal{F}}(n) := \sup_{(\mathbf{x}_1,\dots,\mathbf{x}_n)} |\{(f(\mathbf{x}_1),\dots,f(\mathbf{x}_n)) : f \in \mathcal{F}\}| \tag{1}$$

where $|\cdot|$ denotes the cardinality of the set.

Since the label takes value in $\{-1,1\}$, it follows that $\mathcal{S}_{\mathcal{F}}(n) \leq 2^n$. (Why?) If $\mathcal{S}_{\mathcal{F}}(n) = 2^n$, it means there is a set of n points such that the class of functions $\mathcal{F}$ can generate any classification on these points, in which case we say that $\mathcal{F}$ shatters the set.

a) Show that if $\mathcal{S}_{\mathcal{F}}(N) < 2^N$ for some $N > 1$, then $\mathcal{S}_{\mathcal{F}}(n) < 2^n$ for all $n \geq N$, i.e. that if $\mathcal{F}$ cannot shatter any set of cardinality N, it cannot shatter sets of larger cardinalities.

b) The **VC dimension** of a class $\mathcal{F}$ is defined as the size of the largest set that it can shatter. In other words, the VC dimension of class $\mathcal{F}$ is the largest n such that any assignment of labels can be learned. For a concrete example, let $\mathcal{F} = \left\{f : \mathbb{R}^2 \to \{0,1\} : f \text{ linear classifier}\right\}$ be the function class of linear classifier on the 2-dimensional plane (separating points by a line on the plane).

   i) Compute shattering coefficients $\mathcal{S}_{\mathcal{F}}(1), \mathcal{S}_{\mathcal{F}}(2), \mathcal{S}_{\mathcal{F}}(3)$ and $\mathcal{S}_{\mathcal{F}}(4)$.

   ii) Compute the VC dimension VC-dim($\mathcal{F}$).

c) It turns out that the shattering coefficient can be used as a measure of the 'size' of a class of function as demonstrated by the Vapnik-Chervonenkis theorem:

Theorem 1. ***(Vapnik-Chervonenkis)*** *Let $\mathcal{R}(f), \mathcal{R}_n(f)$ denote the excess risk and empirical risk of the function f resp. For any $\delta > 0$, with probability at least $1-\delta$,*

$$\forall f \in \mathcal{F}, \quad \mathcal{R}(f) - \mathcal{R}_n(f) \leq 2\sqrt{2\frac{\log \mathcal{S}_{\mathcal{F}}(2n) + \log\frac{2}{\delta}}{n}} \tag{2}$$

What condition on the VC dimension of $\mathcal{F}$ is needed for the empirical risk to converge uniformly over the class of functions to the true risk? Explain.

d) In modern maching learning, the models are often over-parameterized and hence the function class $\mathcal{F}$ is enormous. Hence the VC-dimension can be much larger than the size of the training dataset. Show that the bound in Theorem 1 become vacuous when the dataset has size n and $\mathcal{S}_{\mathcal{F}}(2n) = 2^{2n}$.

Further reading: Bousquet et al. (2003) `https://link.springer.com/chapter/10.1007/978-3-540-28650-9_8`
(You can access the article for free if you are within the UniBasel (VPN) network.)

Problem 2 (PAC Bayes Bounds):
Let S denote a training set of size $|S| = m$,

$$S = \{(\mathbf{x}_i, y_i)\}_{i=1,\dots,m}, \text{ where } (\mathbf{x}_i, y_i) \in (\mathcal{X} \times \mathcal{Y}). \tag{3}$$

Let $\mathcal{M}$ denote the set of all probability measures on the data space $\mathbb{R}^k \times \{-1, 1\}$. We will assume that the training examples are i.i.d. samples from some $\mu \in \mathcal{M}$.

A parametric family of classifiers is a function $F : \mathbb{R}^d \times \mathbb{R}^k \to \{-1, 1\}$, where $f_{\mathbf{w}} := F(\mathbf{w}, \cdot) : \mathbb{R}^k \to \{-1, 1\}$ is the classifier indexed by the parameter $\mathbf{w} \in \mathbb{R}^d$. The hypotheses space induced by F is $\mathcal{F} = \{f_{\mathbf{w}} : \mathbf{w} \in \mathbb{R}^d\}$

A randomized classifier is a distribution Q on $\mathbb{R}^d$. Informally, we will speak of distributions on $\mathcal{F}$ when we mean distributions on the underlying parametrization.

We are interested in the 0–1 loss $e_f : \mathbb{R}^k \times \{-1, 1\} \to \{0, 1\}$

$$e_f(\mathbf{x}, y) := \mathbb{I}[f(\mathbf{x}) \neq y].$$

We denote by $e_f^\mu := \mathbb{E}_\mu[e_f(\mathbf{x}, y)]$ the expected error and $e_f^S := \frac{1}{m}\sum_{i=1}^m e_f(\mathbf{x}_i, y_i)$ the empirical error w.r.t. sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$.

The main PAC-Bayesian theorem is the following:

Theorem 2 (McAllister 1999)**.** *For fixed P and any Q, $\epsilon \in (0; 1)$ with prob $\geq 1 - \epsilon$ over sample sets $S \overset{i.i.d.}{\sim} \mu$:*

$$\mathbb{E}_Q[e^\mu_{f_{\mathbf{w}}}] - \mathbb{E}_Q[e^S_{f_{\mathbf{w}}}] \leq \sqrt{\frac{2}{|S|}\left[KL(Q||P) + \ln\left(\frac{2\sqrt{|S|}}{\epsilon}\right)\right]} \tag{4}$$

By reviewing the lecture notes and/or reading the paper by Dziugaite and Roy "*Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data*" `https://arxiv.org/abs/1703.11008`, answer the following questions:

a) What is a "non-vacuous" generalization bound resp. when is a generalization bound considered "non-vacuous" (in contrast to when it is considered to be "vacuous")?

b) In the proof of the McAllister Theorem, we need the condition that Q is absolutely continuous with respect to P. What does this mean and why is it needed?

c) Describe in simple words what the McAllister theorem tells us. What is the role of P and Q? Why is the PAC Bayes bound considered to be a data dependent generalization bound (e.g. in contrast to data independent VC bounds)?

d) How is the PAC Bayes bound used for DNNs in practice? What is a stochastic neural network and how is it implemented in practice?

e) The KL divergence is minimized when $Q = P$. To get bounds on a deterministic classifier, we could imagine choosing Q to be a Dirac-delta distribution which is non-zero for only one weight. Would that work? Explain.

Further reading: Dziugaite and Roy (2017) `https://arxiv.org/abs/1703.11008`

References

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer school on machine learning*, pages 169–207. Springer, 2003.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.