Machine Learning and Neural Networks (MATH3431) Epiphany term, 2024

Handout 3: Learnability, stability

Lecturer & author: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk

Aim. To introduce concepts PAC, fitting vs stability trade off, stability, and their implementation in regularization problems and convex problems.

Reading list & references:

- Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.
  - Ch. 13 Stable rules do not over-fit
- Bousquet, O., Boucheron, S., & Lugosi, G. (2003). Introduction to statistical learning theory. In Summer school on machine learning (pp. 169-207). Berlin, Heidelberg: Springer Berlin Heidelberg. (Suitable for PG students)

1. Learnable Probably Approximately Correct (PAC) Learning

Definition 1. (Agnostic PAC Learnability for General Loss Functions) A hypothesis class $\mathcal{H}$ is agnostic PAC learnable with respect to a set $\mathcal{S}$ and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$, if there exist a function $m_{\mathcal{H}} : (0, 1) \to \mathbb{N}$ and a learning algorithm $\mathfrak{A}(\cdot)$ with the following property: for every $\epsilon \in (0, 1)$, $\delta \in (0, 1)$, and distribution g over $\mathcal{Z}$, when running algorithm $\mathfrak{A}(\cdot)$ given $m \ge m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d examples generated by g then $\mathfrak{A}(\cdot)$ returns $h \in \mathcal{H}$ such as

$$\Pr\left(\hat{R}_{\mathcal{S}}(h) - \min_{h^* \in \mathcal{H}}\left(\hat{R}_{\mathcal{S}}(h^*)\right) \le \epsilon\right) \ge 1 - \delta \tag{1.1}$$

Note 2. It may be easier to work with expectations: by using Markov inequality (1.1) becomes

$$\Pr\left(\hat{R}_{\mathcal{Z}}(h) - \min_{h^* \in \mathcal{H}}\left(\hat{R}_{\mathcal{Z}}(h^*)\right) \le \epsilon\right) \ge 1 - \underbrace{\frac{1}{\epsilon}\mathrm{E}\left(\hat{R}_{\mathcal{Z}}(h) - \min_{h' \in \mathcal{H}}\left(\hat{R}_{\mathcal{Z}}(h')\right)\right)}_{\delta} \tag{1.2}$$

and hence we need to work with expectations and bounded above.

Hint: Markov inequality $\Pr(X \ge a) \le \frac{1}{a}\mathrm{E}(X)$ for $X > 0$.

2. Over-fitting

Note 3. Following we discuss the association of stability and over-fitting. We give an (arguable) definition of over-fitting and based on this we show that: “a learning algorithm does not over-fit if and only if it is on-average-replace-one-stable”.

3. Analysis based on the fitting-stability trade-off

Note 4. Let $R^{*}=\min_{\forall h}\left(R\left(h\right)\right)$ be an ideal/optimal (hence minimum) Risk. The Risk of a learning algorithm $\mathfrak{A}\left(\mathcal{S}\right)$ can be decomposed as

$$R_{g}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-R^{*}=\underbrace{\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-R^{*}}_{\text{(I)}}+\underbrace{R_{g}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)}_{\text{(II)}}\tag{3.1}$$

Note 5. Over-fitting is reasonably quantified by $R_{g}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)$. However, $\hat{R}_{\mathcal{S}}\left(\cdot\right)$ is a random variable and for our computational convenience we focus on expectation w.r.t. $\mathcal{S}$. Hence, we provide the following (arguable) definition of over-fitting (on which we base our analysis).

Definition 6. For a learning algorithm $\mathfrak{A}$, as a measure of over-fitting we consider the expected difference between true and empirical risk

$$\mathrm{E}_{\mathcal{S}\sim g}\left(R_{g}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)\tag{3.2}$$

and we say that $\mathfrak{A}$ suffers from over-fitting when (3.2) is ‘too’ large.

Note 7. The expected Risk of a learning algorithm $\mathfrak{A}\left(\mathcal{S}\right)$ can be written as

$$\mathrm{E}_{\mathcal{S}\sim g}\left(R_{g}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)=\underbrace{\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)}_{\text{(I)}}+\underbrace{\mathrm{E}_{\mathcal{S}\sim g}\left(R_{g}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)}_{\text{(II)}}\tag{3.3}$$

(by applying expectations in (3.1) and ignoring R^{*}), where (I) indicates how well $\mathfrak{A}\left(\mathcal{S}\right)$ fits the training set $\mathcal{D}$, and (II) indicates discrepancy between the true and empirical risks of $\mathfrak{A}\left(\mathcal{S}\right)$. In Section 4, we argue that (II) measures the over-fitting and stability of $\mathfrak{A}\left(\mathcal{S}\right)$.

Note 8. The ultimate goal in a learning problem is to design a learning algorithm $\mathfrak{A}\left(\mathcal{S}\right)$ that both fit the training set and be stable; i.e. minimizing both terms in (3.3).

Note 9. As seen later, there may be a trade-off between (I) and (II); expected empirical risk term and stability term.

4. Stability and over-fitting

Notation 10. Let $\mathcal{S}=\{z_{1},...,z_{m}\}$ be a training sample, and $\mathfrak{A}$ be a learning algorithm with output $\mathfrak{A}\left(\mathcal{S}\right)$.

Note 11. Reasonably a learning algorithm can be stable if a small change of the input to the algorithm does not change the output of the algorithm much. Formalizing this in maths, we can say that if $\mathcal{S}^{(i)}=\{z_{1},...,z_{i-1},z',z_{i+1},...,z_{m}\}$ is another training dataset equal to $\mathcal{S}$ but the ith element which is replaced by another $z'\sim g$, then a good learning algorithm $\mathfrak{A}$ would produce a small value of

$$\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right),z_{i}\right)-\ell\left(\mathfrak{A}\left(\mathcal{S}\right),z_{i}\right)\geq0$$

Definition 12. We say that a learning algorithm $\mathfrak{A}$ is **on-average-replace-one-stable** with rate $\epsilon:\mathbb{N}\rightarrow\mathbb{R}$ if for every distribution g

$$\mathrm{E}\left(\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right),z_{i}\right)-\ell\left(\mathfrak{A}\left(\mathcal{S}\right),z_{i}\right)\right)\leq\epsilon\left(m\right)$$

where $\epsilon(\cdot)$ has to be a decreasing function.

Note 13. Following we discuss the association of stability and over-fitting, based on the over-fitting Definition

Definition 14. [1] A learning algorithm $\mathfrak{A}$ suffers from over-fitting if the expected difference between true and empirical risk

$$\mathrm{E}_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)$$

is large.

Theorem 16. *For any learning algorithm* $\mathfrak{A}$

$$E_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-\hat{R}_{\mathcal{D}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)=E_{\substack{\mathcal{S}\sim g,\, z'\sim g\\ i\sim U\{1,...,m\}}}\left(\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right),z_i\right)-\ell\left(\mathfrak{A}\left(\mathcal{S}\right),z_i\right)\right)$$

where g is a distribution, $\mathcal{S}=\{z_1,...,z_m\}$ and $\mathcal{S}^{(i)}=\{z_1,...,z_{i-1},z',z_{i+1},...,z_m\}$ are *training datasets with $z',z_1,...,z_m\overset{\text{iid}}{\sim}g$.*

Proof. As $z',z_1,...,z_m\overset{\text{iid}}{\sim}g$, then for every i

$$\mathrm{E}_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)=\mathrm{E}_{\substack{\mathcal{S}\sim g\\ z'\sim g}}\left(\ell\left(\mathfrak{A}\left(\mathcal{S}\right),z'\right)\right)=\mathrm{E}_{\substack{\mathcal{S}\sim g\\ z'\sim g}}\left(\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right),z_i\right)\right)$$

and

$$\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)=\mathrm{E}_{\substack{\mathcal{S}\sim g\\ i\sim U\{1,...,m\}}}\left(\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right),z_i\right)\right)$$

□

5. Implementation in regularized loss learning problems

Definition 17. Assume $(\mathcal{H},\mathcal{Z},\ell)$. Regularized Loss Minimization (RLM) learning rule is the one that results as the output of jointly minimizing the empirical risk $\hat{R}_{\mathcal{Z}}(h)$ and a regularization function $J:\mathcal{H}\rightarrow\mathbb{R}$ that is

$$h^{*}=\underbrace{\arg\min}_{h\in\mathcal{H}}\left(\hat{R}_{\mathcal{S}}\left(h\right)+J\left(h\right)\right) \tag{5.1}$$

Remark 18. The motivation for considering the regularization function J in (5.1) is to: (1.) control complexity and (2.) improve stability; as we will see later.

Note 19. We make our example more specific and narrow it to the Ridge RLM learning problem (could be LASSO, Elastic Net, etc.).

1

Note 15. Over-fitting is reasonably quantified by $R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)$. In Definition 14, we use the expectation for our mathematical convenience; other summaries/moments could be used instead.

Definition 20. The Ridge RLM learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$, here $\mathcal{H} = \mathcal{W} \subset \mathbb{R}^d$, uses regularization function $J(w; \lambda) = \lambda \|w\|_2^2$ with $\lambda > 0$, $w \in \mathcal{W}$ and produces learning rule

$$\mathfrak{A}(\mathcal{S}) = \arg\min_{w \in \mathcal{W}} \left(\hat{R}_{\mathcal{S}}(w) + \lambda \|w\|_2^2 \right) \tag{5.2}$$

Note 21. Recall (Term 1) that the regularization function in Ridge RLM learning problem penalizes complexity. Essentially, implies a sequence of hypothesis $\mathcal{H}_1 \subset \mathcal{H}_2 \subset ...$ with $\mathcal{H}_i = \left\{ w \in \mathbb{R}^d : \|w\|_2 < i \right\}$.

Note 22. Below, we will try to analyze the behavior of Ridge RLM learning rule (5.2) w.r.t. the Risk decomposition (3.3). In particular, to upper bounded w.r.t. the the shrinkage term λ,training sample size m, and other characteristics.

5.1. **Bounding the empirical risk (I) in (3.3).**

Note 23. From (5.2), we have

$$\begin{aligned} \hat{R}_{\mathcal{S}}(\mathfrak{A}(\mathcal{S})) &\leq \hat{R}_{\mathcal{S}}(\mathfrak{A}(\mathcal{S})) + \lambda \|\mathfrak{A}(\mathcal{S})\|_2^2 \\ &\leq \hat{R}_{\mathcal{S}}(w') + \lambda \|w'\|_2^2; \quad \forall w' \in \mathcal{W} \end{aligned}$$

and by taking expectations w.r.t. $\mathcal{S}$, it is

$$\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}(\mathfrak{A}(\mathcal{S})) \right) \leq R_g(w') + \lambda \|w'\|_2^2; \quad \forall w' \in \mathcal{W} \tag{5.3}$$

because $\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}(\cdot) \right) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{E}_{\mathcal{S}\sim g}(\ell(\cdot, z_i)) = R_g(\cdot)$.

Note 24. We observe that part (I) in the expected risk decomposition (3.3), (aka the upper bound of the expected empirical risk) increases with the regularization term $\lambda > 0$. (!!!) −Although expected, it did not start well.

5.2. **Bounding the empirical risk (II) in (3.3).**

Note 25. We do that by constraining the loss function to be convex and Lipstchitz.

Assumption 26. *The loss function $\ell(\cdot, z)$ in (5.2) is convex for any $z \in \mathcal{Z}$.*

Note 27. Let $\tilde{R}_{\mathcal{S}}(w) = \hat{R}_{\mathcal{S}}(w) + \lambda \|w\|_2^2$. $\tilde{R}_{\mathcal{S}}(\cdot)$ is 2λ-strongly convex as the sum of a convex function $\hat{R}_{\mathcal{S}}(\cdot)$ (Assumption 26) and a 2λ-strongly convex function $J(\cdot; \lambda) = \lambda \|\cdot\|_2^2$ (results directly from Definition 40).

Note 28. Because $\tilde{R}_{\mathcal{S}}(\cdot)$ is 2λ-strongly convex and $\mathfrak{A}_{\text{Ridge}}(\mathcal{S})$ is its minimizer, according to Lemma 42 in Handout (...), for $\mathfrak{A}(\mathcal{S})$ and any $w \in \mathcal{W}$, it is

$$\tilde{R}_{\mathcal{S}}(w) - \tilde{R}_{\mathcal{S}}(\mathfrak{A}(\mathcal{S})) \geq \lambda \|w - \mathfrak{A}(\mathcal{S})\|^2, \quad \forall w \in \mathcal{W} \tag{5.4}$$

Note 29. Also, for any $w, u \in \mathcal{W}$, it is

$$\begin{aligned}\tilde{R}_{\mathcal{S}}(w) - \tilde{R}_{\mathcal{S}}(u) &= \left(\hat{R}_{\mathcal{S}}(w) + \lambda \|w\|_2^2\right) - \left(\hat{R}_{\mathcal{S}}(u) + \lambda \|u\|_2^2\right) \\ &= \left(\hat{R}_{\mathcal{S}^{(i)}}(w) + \lambda \|w\|_2^2\right) - \left(\hat{R}_{\mathcal{S}^{(i)}} + \lambda \|u\|_2^2\right) \\ &\quad + \frac{\ell(w, z_i) - \ell(u, z_i)}{m} + \frac{\ell(w, z') - \ell(u, z')}{m} \\ &= \tilde{R}_{\mathcal{S}^{(i)}}(w) - \tilde{R}_{\mathcal{S}^{(i)}}(u) + \frac{\ell(w, z_i) - \ell(u, z_i)}{m} + \frac{\ell(w, z') - \ell(u, z')}{m}\end{aligned}$$

Choosing $w = \mathfrak{A}\left(\mathcal{S}^{(i)}\right)$ and $u = \mathfrak{A}(\mathcal{S})$, and the fact that $\tilde{R}_{\mathcal{S}^{(i)}}\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right)\right) \le \tilde{R}_{\mathcal{S}^{(i)}}(\mathfrak{A}(\mathcal{S}))$ it is

$$\tilde{R}_{\mathcal{S}}(w) - \tilde{R}_{\mathcal{S}}(u) \le \frac{\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right), z_i\right) - \ell(\mathfrak{A}(\mathcal{S}), z_i)}{m} + \frac{\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right), z'\right) - \ell(\mathfrak{A}(\mathcal{S}), z')}{m} \tag{5.5}$$

Note 30. Then (5.5) and (5.4) imply

$$\lambda \left\|\tilde{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right)\right) - \tilde{R}_{\mathcal{S}}(\mathfrak{A}(\mathcal{S}))\right\| \le \frac{\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right), z_i\right) - \ell(\mathfrak{A}(\mathcal{S}), z_i)}{m} + \frac{\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right), z'\right) - \ell(\mathfrak{A}(\mathcal{S}), z')}{m} \tag{5.6}$$

Note 31. Now that we brought it in that form, we can use an additional assumption on the loss to bound it.

Assumption 32. *The loss function $\ell(\cdot, z)$ in (5.2) is convex for any $z \in \mathcal{Z}$ and ρ-Lipschitz.*

Note 33. Given ρ-Lipschitzness in Assumption 32, it is

$$\begin{aligned}\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right), z_i\right) - \ell(\mathfrak{A}(\mathcal{S}), z_i) &\le \rho \left\|\mathfrak{A}\left(\mathcal{S}^{(i)}\right) - \mathfrak{A}(\mathcal{S})\right\| \\ \ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right), z'\right) - \ell(\mathfrak{A}(\mathcal{S}), z') &\le \rho \left\|\mathfrak{A}\left(\mathcal{S}^{(i)}\right) - \mathfrak{A}(\mathcal{S})\right\|\end{aligned} \tag{5.7}$$

and hence (5.6) yields

$$\left\|\mathfrak{A}\left(\mathcal{S}^{(i)}\right) - \mathfrak{A}(\mathcal{S})\right\| \le 2\frac{\rho}{\lambda m} \tag{5.8}$$

Note 34. Plugging (5.8) in (5.7) yields

$$\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right), z_i\right) - \ell(\mathfrak{A}(\mathcal{S}), z_i) \le 2\frac{\rho^2}{\lambda m}$$

Note 35. Using Theorem 16, we get an upper bound for the stability / over-fitting

$$\mathrm{E}_{\mathcal{S}\sim g}\left(R_g(\mathfrak{A}(\mathcal{S})) - \hat{R}_{\mathcal{D}}(\mathfrak{A}(\mathcal{S}))\right) = \mathrm{E}_{\substack{\mathcal{S}\sim g,\, z'\sim g \\ i\sim U\{1,\dots,m\}}}\left(\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right), z_i\right) - \ell(\mathfrak{A}(\mathcal{S}), z_i)\right) \le 2\frac{\rho^2}{\lambda m}$$

Note 36. After this saga, the researcher could come to the conclusion that: In a Ridge regularization learning problem with loss function which is convex and ρ-Lipschitz, and the regularizer is $J(\cdot;\lambda) = \lambda\|\cdot\|^2$ with $\lambda > 0$, the learning rule trained against iid sample $\mathcal{S} = \{z_i\}_{i=1}^m$ is on-averave-replace-one-stable with rate $\epsilon(m) = 2\frac{\rho^2}{\lambda m}$; i.e.

$$\mathrm{E}_{\mathcal{S}\sim g}\left(R_g(\mathfrak{A}(\mathcal{S})) - \hat{R}_{\mathcal{D}}(\mathfrak{A}(\mathcal{S}))\right) \le 2\frac{\rho^2}{\lambda m} \tag{5.9}$$

Note 37. From (5.9), we see that stability improves (and over-fitting decreases) as the shrinkage parameter λ increases.

5.3. **Bounding the Risk (3.3).**

Note 38. Given the bounds (5.3) and (5.9), the decomposition of the expected Risk in (3.3) yields that: In a Ridge regularization learning problem with loss function which is convex and ρ-Lipschitz, and the regularizer is $J(\cdot;\lambda) = \lambda \|\cdot\|^2$ with $\lambda > 0$, the learning rule trained against iid sample $\mathcal{S} = \{z_i\}_{i=1}^m$ has expected Risk bound

$$\mathrm{E}_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right) \le \underbrace{R_g\left(w'\right) + \lambda \left\|w'\right\|_2^2}_{\text{(I)}} + \underbrace{2\frac{\rho^2}{\lambda m}}_{\text{(II)}}; \quad \forall w' \in \mathcal{W} \tag{5.10}$$

Note 39. From 5.10, we see that there is a trade-off between Empirical Risk (I) and stability (II) with regards the regularization parameter λ. We wish to use the optimal $\lambda > 0$ corresponding to the smallest bound in (5.10); it has to both fit the training data well (but perhaps not too well) and be very stable to different training data from the same g (but perhaps not too stable)!

Definition 40. (Strongly Convex functions) A function f is λ-strongly convex function is for all w, u, and $\alpha \in (0,1)$ we have

$$f\left(aw + (1-\alpha)u\right) \le af\left(w\right) + (1-\alpha) f\left(u\right) - \frac{\lambda}{2}\alpha\left(1-\alpha\right)\|w-u\|^2 \tag{5.11}$$

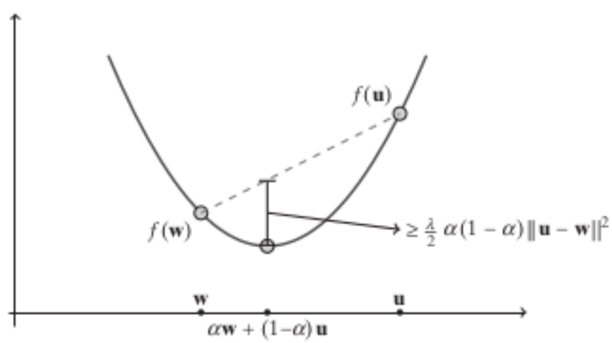

FIGURE 5.1. Strongly convex function

Proposition 41.

1. *The function* $f(w) = \lambda\|w\|_2^2$ *is* 2λ-strongly convex
2. *If* f *is* λ-strongly convex and g is convex then $f+g$ is λ-strongly convex

Lemma 42. *If* f *is* λ-strongly convex and u is a minimizer of f then for any w

$$f(w) - f(u) \ge \frac{\lambda}{2}\|w-u\|^2$$

Proof. Exercise 8 in the Exercise sheet. ∎