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

Handout 3: Learnability and stability in learning problems

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. 2, 3, 13
- 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

Note 1. We formally define the broad learning problem we will work on.

Note 2. Learning algorithms $\mathfrak{A}$ use training data sets $\mathcal{S}$ which may miss characteristics of the unknown data-generating process g. Essentially, approximations (aka errors) are inevitable; some characteristics of data generating process g will be missed even if we use a very representative training set $\mathcal{S}$. We cannot hope that the learning algorithm will find a hypothesis whose error is smaller than the minimal possible error.

Note 3. The PAC learning problem requires no prior assumptions about the data-generating process g. It requires that the learning algorithm $\mathfrak{A}$ will find a predictor $\mathfrak{A}(\mathcal{S})$ whose error is not much larger than the best possible error of a predictor in some given benchmark hypothesis class. So essentially, in practice, the researcher's effort falls on the hypothesis class $\mathcal{H}$.

Definition 4. (Agnostic PAC Learnability for General Loss Functions) A hypothesis class $\mathcal{H}$ is agnostic PAC learnable with respect to a domain $\mathcal{Z}$ of size m and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$, if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm $\mathfrak{A}$ with the following property:

- for every $\epsilon \in (0,1)$, $\delta \in (0,1)$, and distribution g over $\mathcal{Z}$, when running algorithm $\mathfrak{A}$ given training set $\mathcal{S}_m = \{z_1, ..., z_m\}$ with $z_i \overset{\text{IID}}{\sim} g$ for $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ then $\mathfrak{A}$ returns $\mathfrak{A}(\mathcal{S}) \in \mathcal{H}$ such as

$$\Pr_{\mathcal{S} \sim g} \left(R_g\left(\mathfrak{A}\left(\mathcal{S}_m\right)\right) - \min_{h' \in \mathcal{H}} \left(R_g\left(h'\right)\right) \leq \epsilon \right) \geq 1 - \delta \tag{1.1}$$

Note 5. It may be easier to work with expectations: by using Markov inequality (1.1) becomes

$$\Pr_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}_m\right)\right)-\min_{h'\in\mathcal{H}}\left(R_g\left(h'\right)\right)\leq\epsilon\right)\geq 1-\underbrace{\frac{1}{\epsilon}\mathrm{E}_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}_m\right)\right)-\min_{h'\in\mathcal{H}}\left(R_g\left(h'\right)\right)\right)}_{\delta} \tag{1.2}$$

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

Hint: Markov inequality $\Pr\left(X\geq a\right)\leq\frac{1}{a}\mathrm{E}\left(X\right)$ for $X>0$.

Remark 6. The accuracy parameter ϵ determines how far the output rule h can be from the optimal one (this corresponds to the 'approximately correct'). The confidence parameter δ indicates how likely the classifier is to meet that accuracy requirement (corresponds to the "probably" part of "PAC").

2. Analysis of the risk based on the trade-off fitting vs stability

Note 7. Let $R^*=\min_{\forall h}\left(R\left(h\right)\right)$ be an ideal/optimal (hence minimum) Risk, $\mathfrak{A}$ be a learning algorithm, and $\mathfrak{A}\left(\mathcal{S}\right)$ the learnign rule from $\mathfrak{A}$ under training dataset $\mathcal{S}$. The Risk of a learning algorithm $\mathfrak{A}$ can be decomposed as

$$R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-R^*=\underbrace{\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-R^*}_{\text{(I)}}+\underbrace{R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)}_{\text{(II)}} \tag{2.1}$$

Note 8. Over-fitting can be represented 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 here we focus on its expectation w.r.t. $\mathcal{S}\sim g$. Hence, we provide the following (arguable) definition of over-fitting on which we base our analysis .

Definition 9. For a learning algorithm $\mathfrak{A}$, as a measure of over-fitting we consider the expected difference between true Risk and empirical Risk

$$\mathrm{E}_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right) \tag{2.2}$$

Definition 10. We say that learning algorithm $\mathfrak{A}$ suffers from over-fitting when (2.2) is 'too' large.

Note 11. The expected Risk of a learning algorithm $\mathfrak{A}\left(\mathcal{S}\right)$ can be decomposed as

$$\mathrm{E}_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)=\underbrace{\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)}_{\text{(I)}}+\underbrace{\mathrm{E}_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)}_{\text{(II)}} \tag{2.3}$$

(by applying expectations in (2.1) and ignoring R^*), where (I) indicates how well $\mathfrak{A}\left(\mathcal{S}\right)$ fits the training set $\mathcal{S}$, and (II) indicates the discrepancy between the true and empirical risks of $\mathfrak{A}\left(\mathcal{S}\right)$. In Section 3, we argue that (II) measures the over-fitting and a certain type of stability of $\mathfrak{A}\left(\mathcal{S}\right)$.

Note 12. The following result connects the task of upper bounding (and minimizing this bound) the Expected Risk in (2.3) with PAC learning (4).

Theorem 13. *Let $\mathfrak{A}$ be a learning algorithm that guarantees the following:*

- *If* $m \geq m_{\mathcal{H}}(\epsilon)$ *then for every distribution* g*, it is*

$$E_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}_m\right)\right) - \min_{h'\in\mathcal{H}}\left(R_g\left(h'\right)\right)\right) \leq \epsilon$$

Then $\mathfrak{A}$ *it satisfied Definition 4:*

- *for every* $\delta \in (0,1)$*, if* $m \geq m_{\mathcal{H}}(\epsilon\delta)$ then

$$\Pr_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}_m\right)\right) - \min_{h'\in\mathcal{H}}\left(R_g\left(h'\right)\right) \leq \epsilon\right) \geq 1-\delta$$

Proof. Let $\xi = R_g\left(\mathfrak{A}\left(\mathcal{S}_m\right)\right) - \min_{h'\in\mathcal{H}}\left(R_g\left(h'\right)\right)$. From Markov's inequality $\Pr\left(\xi \geq \mathrm{E}\left(\xi\right)/\delta\right) \leq \frac{1}{\mathrm{E}(\xi)/\delta}\mathrm{E}\left(\xi\right) = \delta$. Namely, $\Pr\left(\xi \leq \mathrm{E}\left(\xi\right)/\delta\right) = 1-\delta$. But it is given that if $m \geq m_{\mathcal{H}}\left((\epsilon\delta)\right)$ for every distribution g it is $\mathrm{E}\left(\xi\right) \leq (\epsilon\delta)$. So by substitution $\Pr\left(\xi \leq (\epsilon\delta)/\delta\right) = 1-\delta$ implies $\Pr\left(\xi \leq \epsilon\right) = 1-\delta$. Now substitute back ξ and we conclude the proof. □

Note 14. We aim to design a learning algorithm $\mathfrak{A}\left(\mathcal{S}\right)$ that both fits the training set and is stable; i.e. minimizing both terms in (2.3). As seen later, there may be a trade-off between (I) and (II); expected empirical risk term and stability term. Hence, we aim to upper bound (2.3) by upper bounding (I) and (II) individually.

3. Stability and over-fitting

Notation 15. 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 16. 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

$$\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right), z_i\right) - \ell\left(\mathfrak{A}\left(\mathcal{S}\right), z_i\right) \geq 0$$

Definition 17. We say that a learning algorithm $\mathfrak{A}$ is **on-average-replace-one-stable** with rate $\epsilon : \mathbb{N} \to \mathbb{R}$ if for every distribution g

$$\mathrm{E}_{\substack{\mathcal{S}\sim g \\ z'\sim g, i\sim U\{1,...,m\}}}\left(\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right), z_i\right) - \ell\left(\mathfrak{A}\left(\mathcal{S}\right), z_i\right)\right) \leq \epsilon\left(m\right)$$

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

Note 18. Following we discuss the association of stability and over-fitting, based on the over-fitting Definition 10.

Theorem 19. *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

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

□

4. Implementation in regularized loss learning problems

Definition 20. 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 $J : \mathcal{H} \to \mathbb{R}$ that is

$$h^* = \underbrace{\arg\min}_{h\in\mathcal{H}}\left(\hat{R}_{\mathcal{S}}\left(h\right) + J\left(h\right)\right) \tag{4.1}$$

Remark 21. The motivation for considering the regularization function J in (4.1) is to: (1.) control complexity and (2.) improve stability; as we will see later.

Note 22. We make our example more specific and narrow it to the Ridge RLM learning problem (could be LASSO, Elastic Net, etc.).

Definition 23. The Ridge RLM learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$, here $\mathcal{H} = \mathcal{W} \subset \mathbb{R}^d$, uses regularization function $J\left(w; \lambda\right) = \lambda \left\|w\right\|_2^2$ with $\lambda > 0$, $w \in \mathcal{W}$ and produces learning rule

$$\mathfrak{A}\left(\mathcal{S}\right) = \underset{w\in\mathcal{W}}{\arg\min}\left(\hat{R}_{\mathcal{S}}\left(w\right) + \lambda \left\|w\right\|_2^2\right) \tag{4.2}$$

Note 24. 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 $\mathcal{H}_i = \left\{w \in \mathbb{R}^d : \left\|w\right\|_2 < i\right\}$.

Note 25. Below, we will try to analyze the behavior of Ridge RLM learning rule (4.2) w.r.t. the Risk decomposition (2.3). In particular, to upper bounded w.r.t. the the shrinkage term λ,training sample size m, and other characteristics.

4.1. Bounding the empirical risk (I) in (2.3).

Note 26. From (4.2), we have

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

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

$$\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right) \leq R_g\left(w'\right) + \lambda\left\|w'\right\|_2^2; \quad \forall w' \in \mathcal{W} \tag{4.3}$$

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

Note 27. We observe that part (I) in the expected risk decomposition (2.3), (aka the upper bound of the expected empirical risk) increases with the regularization term $\lambda > 0$. (!!!) –Although expected, it did not start well.

4.2. **Bounding the empirical risk (II) in (2.3).**

Note 28. We do that by constraining the loss function to be convex and Lipstchitz.

Assumption 29. *The loss function $\ell\left(\cdot, z\right)$ in (4.2) is convex for any $z \in \mathcal{Z}$.*

Note 30. Let $\tilde{R}_{\mathcal{S}}\left(w\right) = \hat{R}_{\mathcal{S}}\left(w\right) + \lambda\left\|w\right\|_2^2$. $\tilde{R}_{\mathcal{S}}\left(\cdot\right)$ is 2λ-strongly convex as the sum of a convex function $\hat{R}_{\mathcal{S}}\left(\cdot\right)$ (Assumption 29) and a 2λ-strongly convex function $J\left(\cdot;\lambda\right) = \lambda\left\|\cdot\right\|_2^2$ (results directly from Definition 43).

Note 31. Because $\tilde{R}_{\mathcal{S}}\left(\cdot\right)$ is 2λ-strongly convex and $\mathfrak{A}_{\text{Ridge}}\left(\mathcal{S}\right)$ is its minimizer, according to Lemma 45 in Handout (...), for $\mathfrak{A}\left(\mathcal{S}\right)$ and any $w \in \mathcal{W}$, it is

$$\tilde{R}_{\mathcal{S}}\left(w\right) - \tilde{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right) \geq \lambda\left\|w - \mathfrak{A}\left(\mathcal{S}\right)\right\|^2, \quad \forall w \in \mathcal{W} \tag{4.4}$$

Note 32. Also, for any $w, u \in \mathcal{W}$, it is

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

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

$$\tilde{R}_{\mathcal{S}}\left(w\right) - \tilde{R}_{\mathcal{S}}\left(u\right) \leq \frac{\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right), z_i\right) - \ell\left(\mathfrak{A}\left(\mathcal{S}\right), z_i\right)}{m} + \frac{\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right), z'\right) - \ell\left(\mathfrak{A}\left(\mathcal{S}\right), z'\right)}{m} \tag{4.5}$$

Note 33. Then (4.5) and (4.4) imply

(4.6)

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

Note 34. Now that we brought it in that form, we can use an additional assumption on the loss to bound it.

Assumption 35. *The loss function $\ell\left(\cdot, z\right)$ in (4.2) is convex for any $z \in \mathcal{Z}$ and ρ-Lipschitz.*

Note 36. Given ρ-Lipschitzness in Assumption 35, it is

$$\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right), z_i\right) - \ell\left(\mathfrak{A}\left(\mathcal{S}\right), z_i\right) \leq \rho \left\|\mathfrak{A}\left(\mathcal{S}^{(i)}\right) - \mathfrak{A}\left(\mathcal{S}\right)\right\| \tag{4.7}$$
$$\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right), z'\right) - \ell\left(\mathfrak{A}\left(\mathcal{S}\right), z'\right) \leq \rho \left\|\mathfrak{A}\left(\mathcal{S}^{(i)}\right) - \mathfrak{A}\left(\mathcal{S}\right)\right\|$$

and hence (4.6) yields

$$\left\|\mathfrak{A}\left(\mathcal{S}^{(i)}\right) - \mathfrak{A}\left(\mathcal{S}\right)\right\| \leq 2\frac{\rho}{\lambda m} \tag{4.8}$$

Note 37. Plugging (4.8) in (4.7) yields

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

Note 38. Using Theorem 19, we get an upper bound for the stability / over-fitting

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

Note 39. 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 $J\left(\cdot;\lambda\right) = \lambda\left\|\cdot\right\|^2$ 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 $\epsilon\left(m\right) = 2\frac{\rho^2}{\lambda m}$; i.e.

$$\mathrm{E}_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right) - \hat{R}_{\mathcal{D}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right) \leq 2\frac{\rho^2}{\lambda m} \tag{4.9}$$

Note 40. From (4.9), we see that stability improves (and over-fitting decreases) as the shrinkage parameter λ increases.

4.3. **Bounding the Risk (2.3).**

Note 41. Given the bounds (4.3) and (4.9), the decomposition of the expected Risk in (2.3) yields that: In a Ridge regularization learning problem with loss function which is convex and ρ-Lipschitz, and the regularizer is $J\left(\cdot;\lambda\right) = \lambda\left\|\cdot\right\|^2$ with $\lambda > 0$, the learning rule trained against iid sample $\mathcal{S} = \{z_i\}_{i=1}^m$ has expected Risk bound

$$\mathrm{E}_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right) \leq \underbrace{R_g\left(w'\right) + \lambda\left\|w'\right\|_2^2}_{\text{(I)}} + \underbrace{2\frac{\rho^2}{\lambda m}}_{\text{(II)}}; \quad \forall w' \in \mathcal{W} \tag{4.10}$$

Note 42. From 4.10, 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 (4.10); 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 43. (Strongly Convex functions) A function f is λ-strongly convex function is for all w, u, and $\alpha \in (0,1)$ we have

$$f(aw + (1-\alpha)u) \leq af(w) + (1-\alpha)f(u) - \frac{\lambda}{2}\alpha(1-\alpha)\|w-u\|^2 \tag{4.11}$$

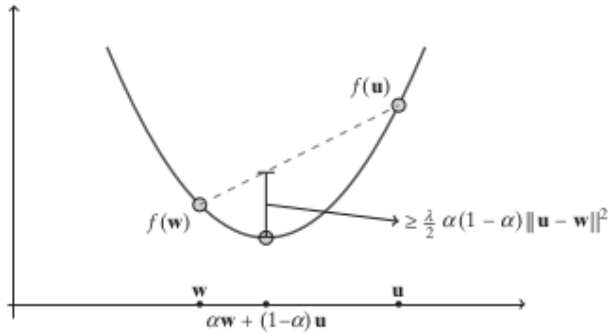

FIGURE 4.1. Strongly convex function

Proposition 44.

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

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

Proof. Exercise 8 in the Exercise sheet. □