



Dong, M., Yang, X., Zhu, R., Wang, Y. and Xue, J.-H. (2020) Generalization Bound of Gradient Descent for Non-Convex Metric Learning. In: 34th Conference on Neural Information Processing Systems (NeurIPS 2020), 06-12 Dec 2020.

This is the Author Accepted Manuscript.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

<http://eprints.gla.ac.uk/225912/>

Deposited on: 6 November 2020

Enlighten – Research publications by members of the University of Glasgow  
<http://eprints.gla.ac.uk>

---

# Generalization Bound of Gradient Descent for Non-Convex Metric Learning

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 Metric learning aims to learn a distance measure that can benefit distance-based  
2 methods such as the nearest neighbor (NN) classifier. While considerable efforts  
3 have been made to improve its empirical performance and analyze its generalization  
4 ability by focusing on the data structure and model complexity, an unresolved ques-  
5 tion is how choices of algorithmic parameters such as training time affect metric  
6 learning as it is typically formulated as an optimization problem and nowadays  
7 more often as a non-convex problem. In this paper, we theoretically address this  
8 question and prove the agnostic Probably Approximately Correct (PAC) learnabil-  
9 ity for metric learning algorithms with non-convex objective functions optimized  
10 via gradient descent (GD); in particular, our theoretical guarantee takes training  
11 time into account. We first show that the generalization PAC bound is a sufficient  
12 condition for agnostic PAC learnability and this bound can be obtained by ensuring  
13 the uniform convergence on a densely concentrated subset of the parameter space.  
14 We then show that, for classifiers optimized via GD, their generalizability can  
15 be guaranteed if the classifier and loss function are both Lipschitz smooth, and  
16 further improved by using fewer iterations. To illustrate and exploit the theoretical  
17 findings, we finally propose a novel metric learning method called *Smooth Metric*  
18 and representative *Instance LEarning* (SMILE), designed to satisfy the Lipschitz  
19 smoothness property and learned via GD with an early stopping mechanism for  
20 better discriminability and less computational cost of NN.

## 21 1 Introduction

22 A good measure of distance between instances is important to many machine learning algorithms,  
23 such as the nearest neighbor (NN) classifier and  $k$ -means clustering. As it is difficult to handcraft  
24 an optimal distance for each task, metric learning appears as an appealing technique to learn the  
25 distance metric automatically and directly from the data. The most widely studied metric is the  
26 Mahalanobis distance and it is often learned as an optimization problem [46, 16, 44]. To enhance the  
27 discriminability of the learned metric, various loss functions have been designed, considering the local  
28 property of heterogeneous data [14, 42, 21, 4, 33, 49, 39, 11] and the nonlinear geometry of the sample  
29 space [22, 53, 7]. Meanwhile, to achieve good generalization and robustness, different regularizations  
30 have been imposed to control the model complexity [27, 24, 45, and references therein]. In addition to  
31 methodological advances, theoretical guarantees of metric learning algorithms, as well as guarantees  
32 of metric-based classifiers [2, 18], have been provided. In particular, generalization bounds have  
33 been founded on the complexity measure of the model class [50, 3, 6, 41, 29, 48], algorithmic  
34 stability [25, 18, 15], and algorithmic robustness [1]. The intrinsic complexity of the dataset has also  
35 been considered in recent works [41, 29].

36 While the data structure and model complexity play a vital role in metric learning, an equally  
37 important but as yet poorly understood factor is the choice of optimization algorithms and the  
38 associated parameters [37]. For example, when metric learning is formulated as a non-convex  
39 problem and optimized by using the gradient descent algorithm, its solution is inevitably influenced  
40 by factors such as the learning rate and the number of training iterations; different local minima will  
41 then exhibit different generalization behavior.

42 Therefore, the goal of this paper is to provide a new route to theoretical exploration and exploitation  
43 of the effect of the gradient descent (GD) algorithm on metric learning methods. To this end, we  
44 provide a generalization bound which suggests that early stopping, smooth classifier and smooth loss  
45 function have crucial influence on the generalization error. We highlight that our results are obtained  
46 without using any property of convex optimization, and hence are applicable to non-convex metric  
47 learning methods. The contributions of this paper are fourfold.

48 1. We show that the generalization Probably Approximately Correct (PAC) bound, which is a weaker  
49 notion than the uniform convergence condition, is a sufficient condition for a parametric hypothesis  
50 class to be agnostic PAC learnable (Theorem 1).

51 2. To facilitate the derivation of the generalization PAC bound of a hypothesis class, we propose  
52 a new decomposition theorem to decompose the bound into two terms that can be easily guaran-  
53 teed (Theorem 2). The first term constrains the space of the estimated parameters of the hypothesis,  
54 reducing it from the entire parameter space to a high-confidence subset of the parameter space. The  
55 second term considers the uniform convergence condition of the concentrated subset.

56 3. Based on the decomposition theorem, we obtain the generalization PAC bound for classifiers  
57 learned with the gradient descent algorithm (Theorem 3). The bound shows that the generalization gap  
58 increases over iterations, thus providing a theoretical support for the practical use of early stopping.  
59 Moreover, it shows that a Lipschitz smooth (i.e. Lipschitz continuous of the gradient) classifier and a  
60 Lipschitz smooth loss function are necessary for generalization guarantee.

61 4. We propose a novel metric learning method as a concrete example of using the generalization PAC  
62 bound. When classifying a test instance, the NN classifier has to store the entire training set and  
63 calculate its distances to all training instances, thereby incurring high storage and computational costs.  
64 To reduce these costs and improve the generalization performance, we propose to simultaneously  
65 learn the distance metric and few representative instances which serve as the reference points for  
66 testing; the new method is called *Smooth Metric and representative Instance LEarning* (SMILE).  
67 More specific, to ensure good test performance, SMILE adopts a Lipschitz smooth classifier and  
68 loss function and is optimized via GD with a designed early stopping mechanism. The method is  
69 evaluated on 12 datasets and shows competitive performance against existing methods.

## 70 1.1 Related work

71 **Generalization bound of GD with early stopping** Early stopping in regularizing the model com-  
72 plexity and its effect on the generalization ability have been extensively studied for a wide range of  
73 methods, such as perceptron algorithm [8], kernel regression [47], and deep neural networks [31]. Our  
74 algorithm-dependent PAC bound is motivated by [20], which proves the generalizability for models  
75 learned with stochastic GD. The main difference between [20] and our work is that [20] studies  
76 the expected generalization gap, which is not a sufficient condition for agnostic PAC learnability,  
77 whereas the generalization PAC bound studied in this paper is a sufficient condition. Consequently,  
78 we need a new decomposition theorem so that the generalization PAC bound can be used to analyze  
79 models learned with GD.

80 **Generalization bounds for Lipschitz classifiers and losses** [28, 17] use Lipschitz functions as  
81 large margin classifiers in general metric spaces and provide generalization bounds for Lipschitz  
82 classifiers. Our theoretical guarantee is different from their work in two aspects. First, the input  
83 space of the Lipschitz constant is the data space in [28, 17], whereas the input space is the parameter  
84 space in our paper. Second, owing to this difference, the generalization bound obtained in our work  
85 has a faster convergence in most cases. [41] derives the generalization bound for metric learning  
86 algorithms with Lipschitz continuous loss functions. However, when taking the influence of GD into  
87 account, Lipschitz continuity is not sufficient to guarantee the generalizability; Lipschitz smoothness  
88 is also needed. [48] makes use of a smooth loss function to obtain a fast generalization. However,

89 their work requires the objective function to be strongly convex, which is different from our focus on  
 90 non-convex problems.

91 **Metric learning with representative instances** Reducing the amount of necessary training data  
 92 as a way of reducing the storage and computational costs of NN has been extensively studied, e.g.  
 93 in [30, 52, 35]. Among these methods, SNC [26] and ProtoNN [19] are the most relevant to our work,  
 94 as they also learn the distance metric and representative instances simultaneously. Our method differs  
 95 from them in the loss function and regularization terms, both of which are designed in our work to  
 96 provide a theoretical guarantee on the classification performance.

## 97 2 Preliminaries

### 98 2.1 Notation

99 This paper focuses on binary classification problems. Let  $\mathbf{z}^n = \{\mathbf{z}_i = (\mathbf{x}_i, y_i), i = 1, \dots, n\} \in \mathcal{Z}^n$   
 100 denote the set of  $n$  independent and identically distributed (i.i.d.) training instance and label pairs,  
 101 sampled from an unknown joint distribution  $p(\mathbf{z}) = p(\mathbf{x}, y)$ . Let  $h(\mathbf{x}, \mathbf{w})$  be a function with instance  
 102  $\mathbf{x}$  and parameter  $\mathbf{w} \in \mathcal{W} \subseteq \mathbb{R}^Q$ . The output of  $h(\mathbf{x}, \mathbf{w})$  is restricted to be a real value;  $\text{sign}[h(\mathbf{x}, \mathbf{w})]$   
 103 returns the classification decision, where  $\text{sign}[\cdot]$  denotes the sign function.

104 During the training process of the classifier, given  $\mathbf{z}^n$ , a *classifier* or *hypothesis*  $\hat{h}$  can be obtained  
 105 from an optimization algorithm, such as GD.  $R_n(\mathbf{z}^n, \hat{h}) := \frac{1}{n} \sum_i r(\mathbf{z}_i, \hat{h}) := \frac{1}{n} \sum_i l(\hat{h}(\mathbf{x}_i), y_i)$  is  
 106 called the *training error*, where  $r(\cdot, \cdot)$  denotes the risk function and  $l(\cdot, \cdot)$  denotes the loss function.  
 107 Let  $\mathbf{s} \in \mathcal{S}$  denote a fixed setting of the algorithm, including e.g. the initial values, the number  
 108 of iterations and the learning rate. With a parametric classifier,  $\hat{h}$  can be fully represented by  $\hat{\mathbf{w}}$ .  
 109 The relationship between  $\mathbf{w}$  and  $\mathbf{z}^n$  is represented as  $\hat{\mathbf{w}} = \mathbf{m}(\mathbf{z}^n; \mathbf{s})$ , where  $\mathbf{m} : \mathcal{Z}^n \times \mathcal{S} \rightarrow \mathcal{W}$ ;  
 110  $\mathbf{m}(\mathbf{z}^n; \mathbf{s})$  will sometimes be abbreviated to  $\mathbf{m}(\mathbf{z}^n)$  for notational simplicity. Since  $\hat{\mathbf{w}}$  is a function  
 111 of random samples  $\mathbf{z}^n$ ,  $\hat{\mathbf{w}}$  is also a random variable. In the parametric case, the training error will be  
 112 represented as  $R_n(\mathbf{z}^n, \hat{\mathbf{w}}) := \frac{1}{n} \sum_i r(\mathbf{z}_i, \hat{\mathbf{w}}) := \frac{1}{n} \sum_i l(h(\mathbf{x}_i, \hat{\mathbf{w}}), y_i)$ .

113 During the test process, a test pair  $\mathbf{z}' = (\mathbf{x}', y')$  is sampled from the same unknown distribution  $p(\mathbf{z})$ .  
 114 The predicted value  $h(\mathbf{x}', \hat{h})$  will be compared with the true label  $y'$  to evaluate the performance of  
 115 the algorithm.  $R(\hat{h}) := \mathbb{E}_{\mathbf{z}'} r(\mathbf{z}', \hat{h}) := \mathbb{E}_{\mathbf{z}'} l(h(\mathbf{x}', \hat{h}), y')$  is called the *test error*. With a parametric  
 116 classifier, the following notations will be used  $R(\hat{\mathbf{w}}) := \mathbb{E}_{\mathbf{z}'} r(\mathbf{z}', \hat{\mathbf{w}}) := \mathbb{E}_{\mathbf{z}'} l(h(\mathbf{x}', \hat{\mathbf{w}}), y')$ .

117 The gap between the training error and the test error,  $R(\hat{\mathbf{w}}) - R_n(\mathbf{z}^n, \hat{\mathbf{w}})$ , is called the *generalization*  
 118 *gap*. A good classifier should have small training error and small generalization gap so as to perform  
 119 well on test instances.

120 Let  $\|\mathbf{a}\|_2$  denote the  $L_2$ -norm of a vector  $\mathbf{a}$  and  $\|\mathbf{A}\|_F$  denote the Frobenius norm of a matrix  $\mathbf{A}$ .  
 121 The subscript of norm will be dropped when it is clear from the context.  $\mathbf{a}_{[q]}$  denotes the  $q$ th element  
 122 of a vector  $\mathbf{a}$ .

### 123 2.2 Definitions

124 **Definition 1.** [43] Let  $(\Theta, \rho_\Theta)$ ,  $(\mathcal{V}, \rho_\mathcal{V})$  be two metric spaces. A function  $h : \Theta \rightarrow \mathcal{V}$  is called  
 125 *Lipschitz continuous* if  $\exists L < \infty, \forall \theta_1, \theta_2 \in \Theta$ ,

$$\rho_\mathcal{V}(h(\theta_1), h(\theta_2)) \leq L \rho_\Theta(\theta_1, \theta_2).$$

126 The *Lipschitz constant* of  $h$  with respect to the input space  $\Theta$ , denoted by  $\text{lip}(h; \mathcal{V} \leftarrow \Theta)$  or  
 127  $\text{lip}(h \leftarrow \Theta)$  for short, is the smallest  $L$  such that the above inequality holds.

128 **Definition 2.** A function  $r : \Theta \rightarrow \mathbb{R}$  is called *Lipschitz smooth*, if  $\exists \eta < \infty, \forall \theta_1, \theta_2 \in \Theta$ ,

$$\|\nabla r(\theta_1) - \nabla r(\theta_2)\| \leq \eta \|\theta_1 - \theta_2\|.$$

129 The Lipschitz constant of the derivative of  $r$  with respect to  $\Theta$ , denoted by  $\text{lip}(\frac{\partial r}{\partial \theta} \leftarrow \Theta)$ , is the  
 130 smallest  $\eta$  such that the above inequality holds.

131 Some properties of Lipschitz functions are frequently used in the paper, such as constructing sophisti-  
 132 cated Lipschitz functions from the basic ones and bounding the Lipschitz constant via the gradient of  
 133 differentiable functions; details are listed in Appendix A.

134 **Definition 3.** [43] The *diameter* of a set  $\mathcal{V}$  is defined as

$$\text{diam}(\mathcal{V}) = \max_{\mathbf{v}_i, \mathbf{v}_j \in \mathcal{V}} \|\mathbf{v}_i - \mathbf{v}_j\|.$$

### 135 3 Learnability via the generalization PAC Bound

136 In this section, we first introduce the generalization PAC bound and establish its link with the agnostic  
 137 PAC learnability. We then propose a decomposition theorem. Finally, we apply the theorem to prove  
 138 the learnability of the gradient descent algorithm.

#### 139 3.1 Generalization PAC bound and agnostic PAC learnability

140 One classical way of determining whether a hypothesis class is agnostic PAC learnable is to verify  
 141 the uniform convergence condition, which bounds the generalization gap over all hypotheses of  
 142 the class. However, as some hypotheses are not searched under a fixed setting of the optimization  
 143 algorithm, [5] proposes to bound the generalization gap for specific algorithms. We adopt this notion  
 144 and formally define the generalization PAC bound as follows.

145 **Definition 4.** A hypothesis class  $\mathcal{H}$  has the *generalization PAC bound* if there exists a function  
 146  $n_{\mathcal{H}}^G : (0, 1)^2 \rightarrow \mathbb{N}$  such that for every  $\epsilon, \delta \in (0, 1)$  and for every probability distribution  $\mathcal{D}_{\mathcal{Z}}$  over  $\mathcal{Z}$ ,  
 147 if  $\mathbf{z}^n$  is a sample of  $n \geq n_{\mathcal{H}}^G(\epsilon, \delta)$  i.i.d. examples drawn from  $\mathcal{D}_{\mathcal{Z}}$ , the algorithm returns a hypothesis  
 148  $\hat{h}$  such that the following inequality is satisfied:

$$\mathbb{P}_{\mathbf{z}^n}[R(\hat{h}) - R_n(\mathbf{z}^n, \hat{h}) \leq \epsilon] \geq 1 - \delta. \quad (1)$$

149 First, we note that  $\hat{h}$  is regarded as a random variable in this paper. Second, while the generalization  
 150 PAC bound is a weaker condition than the uniform convergence, as shown in Lemma 1, it is still a  
 151 sufficient condition for the agnostic PAC learnability, as shown in Theorem 1. Proofs of theorems  
 152 and lemmas are given in Appendix C.

153 **Lemma 1.** The relationship between the generalization PAC bound and the uniform convergence  
 154 bound is as follows:

$$\mathbb{P}_{\mathbf{z}^n}[R(\hat{h}) - R_n(\mathbf{z}^n, \hat{h}) \leq \epsilon] \geq \mathbb{P}_{\mathbf{z}^n}\left[\max_{h \in \mathcal{H}} (R(h) - R_n(\mathbf{z}^n, h)) \leq \epsilon\right]. \quad (2)$$

155 **Theorem 1.** Suppose  $\text{ERM}_{\mathcal{H}}$  exists for a class  $\mathcal{H}$ , where  $\text{ERM}_{\mathcal{H}}$  denotes the empirical risk  
 156 minimization learner over the class  $\mathcal{H}$ . If  $\mathcal{H}$  has the generalization PAC bound with a function  
 157  $n_{\mathcal{H}}^G : (0, 1)^2 \rightarrow \mathbb{N}$ , then  $\mathcal{H}$  is agnostic PAC learnable with the sample complexity function  
 158  $n_{\mathcal{H}}^{AL}(\epsilon, \delta) \leq \max[n_{\mathcal{H}}^G(\epsilon/2, \delta/2), \frac{2C_r^2}{\epsilon^2} \ln \frac{4}{\delta}]$ , where the range of the risk function  $r(\mathbf{z}, h)$  is  $[0, C_r]$ .  
 159 Furthermore, in this case,  $\text{ERM}_{\mathcal{H}}$  is a successful agnostic PAC learner for  $\mathcal{H}$ .

#### 160 3.2 Decomposition theorem for the generalization PAC bound

161 Directly bounding Eq. 1 is difficult due to the random nature of  $\mathbf{z}^n$  and  $\hat{h}$  in  $R_n$ . To disentangle these  
 162 two quantities, we propose the following decomposition theorem. Its core idea is to use the uniform  
 163 convergence bound in a much smaller set.

164 **Theorem 2.** (Decomposition Theorem) Let  $\mathcal{W}$  denote the set of all possible values of  $\mathbf{w}$  and  $\hat{\mathcal{W}} \subseteq \mathcal{W}$ ;  
 165 let  $\delta_1, \delta_2 \geq 0$ . If

$$\mathbb{P}_{\mathbf{z}^n}[\hat{\mathbf{w}} \in \hat{\mathcal{W}}] \geq 1 - \delta_1 \quad (3)$$

166 and

$$\mathbb{P}_{\mathbf{z}^n}\left[\max_{\mathbf{w} \in \hat{\mathcal{W}}} (R(\mathbf{w}) - R_n(\mathbf{z}^n, \mathbf{w})) \leq \epsilon\right] \geq 1 - \delta_2, \quad (4)$$

167 then the following inequality holds:

$$\mathbb{P}_{\mathbf{z}^n}[R(\hat{\mathbf{w}}) - R_n(\mathbf{z}^n, \hat{\mathbf{w}}) \leq \epsilon] \geq 1 - \delta_1 - \delta_2. \quad (5)$$

168 Theorem 2 decomposes the generalization PAC bound into two terms which are easier to be controlled,  
 169 namely (i) a smaller parameter space  $\hat{\mathcal{W}}$  that includes estimated parameter vectors with  
 170 high probability; (ii) uniform convergence of  $\hat{\mathcal{W}}$ . In the following section, the theorem is applied to  
 171 analyze the generalization ability of the gradient descent algorithm. We show that term (i) can be  
 172 bounded by applying the concentration inequality to the random variables  $\hat{\mathbf{w}}$  and term (ii) can be  
 173 bounded based on the covering number.

174 **3.3 Learnability of the gradient descent algorithm**

175 **3.3.1 Settings**

176 The updating equation of the most conventional GD algorithm is as follows:

$$\begin{aligned}
 \hat{\mathbf{w}}^{(1)} &= \mathbf{w}^{(0)} - \frac{\alpha^{(1)}}{n} \sum_{i=1}^n \frac{\partial r(\mathbf{z}_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\hat{\mathbf{w}}^{(0)}}; \\
 &\vdots \\
 \hat{\mathbf{w}}^{(T)} &= \hat{\mathbf{w}}^{(T-1)} - \frac{\alpha^{(T)}}{n} \sum_{i=1}^n \frac{\partial r(\mathbf{z}_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\hat{\mathbf{w}}^{(T-1)}} \\
 &= \mathbf{w}^{(0)} - \sum_{t=1}^T \frac{\alpha^{(t)}}{n} \sum_{i=1}^n \frac{\partial r(\mathbf{z}_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\hat{\mathbf{w}}^{(t-1)}},
 \end{aligned}$$

179 where  $\alpha^{(t)} \geq 0$  denotes the learning rate at iteration  $t$ ;  $\hat{\mathbf{w}}^{(t)}$  denotes the estimated parameters of  
 180 the classifier obtained after  $t$  iterations;  $\mathbf{w}^{(0)}$  denotes the initial parameter of the algorithm. Here  
 181 the number of iterations  $T$  and the learning rate  $\alpha^{(t)}$  are treated as the setting parameters of the GD  
 182 algorithm and determined in advance, i.e.  $\mathbf{s} = \{T, \alpha^{(t)}, t = 1, \dots, T\}$ . The initial weight  $\mathbf{w}^{(0)}$   
 183 is assumed to be fixed.

184 **3.3.2 Concentration of  $\hat{\mathbf{w}}^{(T)}$**

185 Recall that  $\mathbf{m}^{(T)}(\mathbf{z}^n; \mathbf{s}) = \hat{\mathbf{w}}^{(T)} \in \mathbb{R}^Q$  and  $\mathbf{m}_{[q]}^{(T)}(\mathbf{z}^n; \mathbf{s})$  denotes the  $q$ th element of  $\mathbf{m}^{(T)}(\mathbf{z}^n; \mathbf{s})$ .  
 186 To prove that the first term of Theorem 2 holds, we set  $\hat{\mathcal{W}}$  as the Euclidean ball centered at  
 187  $\mathbb{E}_{\mathbf{z}^n} \mathbf{m}(\mathbf{z}^n; \mathbf{s})$  with radius  $\epsilon$ , denoted by  $\text{ball}(\mathbb{E}_{\mathbf{z}^n} \mathbf{m}(\mathbf{z}^n; \mathbf{s}), \epsilon)$ . The condition that  $\hat{\mathbf{w}} \in \hat{\mathcal{W}}$  with  
 188 high probability is equivalent to the condition that  $\mathbf{m}(\mathbf{z}^n; \mathbf{s}) \in \text{ball}(\mathbb{E}_{\mathbf{z}^n} \mathbf{m}(\mathbf{z}^n; \mathbf{s}), \epsilon)$  with high  
 189 probability. With a fixed setting  $\mathbf{s}$  and any fixed initial parameter vector  $\mathbf{w}^{(0)}$ , given the training  
 190 samples  $\mathbf{z}^n$ , the value of  $\mathbf{m}_{[q]}^{(T)}(\mathbf{z}^n; \mathbf{s})$  is determined. In other words,  $\mathbf{m}_{[q]}^{(T)}(\mathbf{z}^n; \mathbf{s})$  is a function from  
 191  $\mathcal{Z}^n$  to  $\mathbb{R}$ . By applying the McDiarmid's inequality (Lemma B.1), we obtain the following lemma on  
 192 the concentration property of  $\mathbf{m}^{(T)}(\mathbf{z}^n; \mathbf{s})$ .

193 **Lemma 2.** The following bound holds for any fixed  $\mathbf{s}$  and  $\mathbf{w}^{(0)}$ :

$$\mathbb{P}_{\mathbf{z}^n} \left[ \mathbf{m}^{(T)}(\mathbf{z}^n; \mathbf{s}) \in \text{ball}(\mathbb{E}_{\mathbf{z}^n} \mathbf{m}^{(T)}(\mathbf{z}^n; \mathbf{s}), \epsilon) \right] \geq 1 - 2Q \exp\left(\frac{-2\epsilon^2 n}{QC^2}\right), \quad (6)$$

194 where  $\mathbf{m}^{(T)}(\mathbf{z}^n; \mathbf{s}) \in \mathbb{R}^Q$ ;  $C = 2(\sum_{t=1}^T \eta^{T-t} \alpha^{(t)}) \text{lip}(r \leftarrow \mathbf{w})$ ;  $\eta = \text{lip}(G \leftarrow \mathbf{w})$  and  
 195  $G(\mathbf{m}^{(t-1)}(\mathbf{z}^n)) = \mathbf{m}^{(t-1)}(\mathbf{z}^n) - \frac{\alpha^{(t)}}{n} \sum_{j \in [n]/i} \frac{\partial r(\mathbf{z}_j, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{m}^{(t-1)}(\mathbf{z}^n)}$ ;  $[n]/i$  denotes the set which  
 196 contains the integers from 1 to  $n$  without  $i$ .

197 The key idea behind the proof is as follows. Randomness of sampling leads to randomness of the  
 198 learned parameter vector  $\hat{\mathbf{w}}$ . After one iteration of gradient update, the difference between  $\hat{\mathbf{w}}$  learned  
 199 on the random samples and that learned on the population is controlled by the Lipschitz constant of  $r$   
 200 and  $G$ . Such differences will accumulate over iterations, thereby affecting the concentration property.

201 **3.3.3 Uniform convergence inside  $\hat{\mathcal{W}}$**

202 The following uniform convergence condition is obtained based on the covering number and Dudley's  
 203 chaining integral [13]. By using the Lipschitz constant, we can bound the covering number of the  
 204 hypothesis class by the covering number of the parameter space.

205 **Lemma 3.** Suppose  $\text{lip}(r \leftarrow \mathbf{w}) \leq L$  and  $\text{diam}(\mathcal{W}) \leq B$ , then the following inequality holds:

$$\mathbb{P}_{\mathbf{z}^n} \left[ \max_{\mathbf{w} \in \mathcal{W}} (R(\mathbf{w}) - R_n(\mathbf{z}^n, \mathbf{w})) \leq CLB \sqrt{\frac{Q}{n}} + \sqrt{\frac{\ln(1/\delta)}{2n}} \right] \geq 1 - \delta, \quad (7)$$

206 where  $C$  is a universal constant.

207 **3.3.4 Application of the decomposition theorem**

208 **Theorem 3.** Suppose  $\text{lip}(h \leftarrow \mathbf{w}) \leq L_1$  and  $\text{lip}(r \leftarrow h) \leq L_l$ . Then with probability at least  
209  $1 - \delta_1 - \delta_2$ , the following bound holds:

$$R(\mathbf{m}(\mathbf{z}^n; \mathbf{s})) - R_n(\mathbf{z}^n, \mathbf{m}(\mathbf{z}^n; \mathbf{s})) \leq \frac{C_1 C_2 L_1^2 L_l^2 Q \sqrt{\ln(2Q/\delta_1)}}{n} + \sqrt{\frac{\ln(1/\delta_2)}{2n}}, \quad (8)$$

210 where  $\mathbf{w} \in \mathbb{R}^Q$ ;  $C_1$  is a universal constant;  $C_2 = \sum_{t=1}^T \eta^{T-t} \alpha^{(t)}$ , in which  $T$  denotes the number  
211 of iterations,  $\alpha^{(t)}$  denotes the learning rate at time  $t$ ,  $\eta = \text{lip}(G \leftarrow \mathbf{w})$ , and  $G(\mathbf{m}^{(t-1)}(\mathbf{z}^n)) =$   
212  $\mathbf{m}^{(t-1)}(\mathbf{z}^n) - \frac{\alpha^{(t)}}{n} \sum_{j \in [n]/i} \frac{\partial r(\mathbf{z}_j, \mathbf{w})}{\partial \mathbf{w}} |_{\mathbf{m}^{(t-1)}(\mathbf{z}^n)}$ ;  $[n]/i$  denotes the set which contains the integers  
213 from 1 to  $n$  without  $i$ .

214 Theorem 3 suggests that the following factors will affect the generalizability of the learned model.  
215 1)  $T$ : A smaller number of iterations leads to better concentration property and thus better generaliza-  
216 tion performance. Thus, when optimizing via GD, we select the model from the earliest iteration  $t$   
217 that yields the minimum training error; the test stage is implemented using the parameters learned  
218 at  $t$ ;  
219 2)  $Q$ : A smaller value of  $Q$ , i.e. fewer parameters, gives a tighter generalization bound;  
220 3)  $L_1, L_l$ : Using a classifier and loss function with smaller Lipschitz constants will improve the  
221 generalizability;  
222 4)  $\eta$ : Based on the definition of  $G$  and the addition property of Lipschitz functions (Appendix A),  
223 if  $\text{lip}(\frac{\partial r(\mathbf{z}_j, \mathbf{w})}{\partial \mathbf{w}} \leftarrow \mathbf{w})$  is bounded by  $L_s$ , then  $\eta$  is bounded by  $1 + \alpha L_s$ . Based on the composition  
224 property of Lipschitz functions, we have

$$\text{lip}\left(\frac{\partial r}{\partial \mathbf{w}} \leftarrow \mathbf{w}\right) = \text{lip}\left(\frac{\partial r}{\partial h} \frac{\partial h}{\partial \mathbf{w}} \leftarrow \mathbf{w}\right) \leq \text{lip}\left(\frac{\partial r}{\partial h} \leftarrow \mathbf{w}\right) \text{lip}\left(\frac{\partial h}{\partial \mathbf{w}} \leftarrow \mathbf{w}\right).$$

225 Thus  $\eta$  is bounded if both  $\text{lip}(\frac{\partial h}{\partial \mathbf{w}} \leftarrow \mathbf{w})$  and  $\text{lip}(\frac{\partial r}{\partial h} \leftarrow h)$  are bounded. In other words, the classifier  
226 and loss function should be Lipschitz smooth.

227 **4 Smooth metric and representative instance learning (SMILE)**

228 Theorem 3 shows that Lipschitz smoothness is indispensable for ensuring generalization. To enjoy  
229 and illustrate the practical exploitation of this appealing theoretical result, we establish a simple yet  
230 theoretically well-founded and new metric learning method called SMILE with a smooth classifier  
231 and a smooth loss function. SMILE learns a Mahalanobis distance to enhance the classification  
232 performance of NN classifier. Meanwhile, to reduce the storage and computational cost of NN,  
233 SMILE learns few representative instances in the training stage and calculate the distances between  
234 the test instance and representative instances only in the test stage. In this section, we present the  
235 classifier, the loss function, the optimization problem, and some experimental results of SMILE.

236 **4.1 The classifier of SMILE**

237 For any two instances  $\mathbf{x}_i$  and  $\mathbf{x}_j$ , the generalized (squared) Mahalanobis distance is defined as  
238  $d_M^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)$ , where  $\mathbf{M}$  is a positive semidefinite (PSD) matrix. Owing  
239 to the PSD property,  $\mathbf{M} = \mathbf{L}^T \mathbf{L}$  and thus  $d_M^2(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{L}\mathbf{x}_i, \mathbf{L}\mathbf{x}_j) = \|\mathbf{L}\mathbf{x}_i - \mathbf{L}\mathbf{x}_j\|_2^2$ .

240 The classifier of SMILE is simply defined as follows:

$$h(\mathbf{x}; \mathbf{r}^m, \mathbf{L}) = \sum_j \exp(-d^2(\mathbf{L}\mathbf{x}, \mathbf{r}_j^+)) - \sum_k \exp(-d^2(\mathbf{L}\mathbf{x}, \mathbf{r}_k^-)), \quad (9)$$

241 where  $\mathbf{r}^m$  and  $\mathbf{L}$  are the parameters of the classifier;  $\mathbf{r}_j^+$  and  $\mathbf{r}_k^-$  denote the  $j$ th representative instance  
242 of the positive class and the  $k$ th representative instance of the negative class, respectively;  $m$  denotes  
243 the total number of learned representative instances. The test instance  $\mathbf{x}$  is classified to the positive  
244 class when  $h(\mathbf{x}) \geq 0$  and to the negative class when  $h(\mathbf{x}) < 0$ .

245 As shown in Appendix D, a sufficient condition for  $h$  to be Lipschitz smooth is that  $\text{diam}(\mathbf{L})$ ,  
246  $\text{diam}(\mathbf{x})$  and  $\text{diam}(\mathbf{r})$  are bounded. With a slight abuse of notation,  $\text{diam}(\mathbf{L})$  denotes the diameter  
247 of the set which contains all possible values of  $\mathbf{L}$ ;  $\text{diam}(\mathbf{x})$  and  $\text{diam}(\mathbf{r})$  are defined similarly. To  
248 bound these quantities, we will constrain the Frobenius norm of  $\mathbf{L}$  and the  $L_2$ -norm of  $\mathbf{x}$  and  $\mathbf{r}$ .

249 **4.2 The loss function of SMILE**

250 Similarly to the Huber loss for regression [23], we propose the following loss function defined by  
 251 combining a quadratic and a linear function:

$$l(a) = \begin{cases} 1 - a & \text{if } a \leq 0; \\ \frac{1}{4}(a - 2)^2 & \text{if } 0 < a \leq 2; \\ 0 & \text{if } a > 2. \end{cases} \quad (10)$$

The derivative of  $l(a)$  is as follows:

252 
$$l'(a) = \begin{cases} -1 & \text{if } a \leq 0; \\ \frac{a-2}{2} & \text{if } 0 < a \leq 2; \\ 0 & \text{if } a > 2. \end{cases}$$

The loss function and its derivative are illustrated in Figure 1. The Lipschitz constant of  $l'(a)$  is  $\frac{1}{2}$ , meaning that the proposed loss is a Lipschitz smooth function.

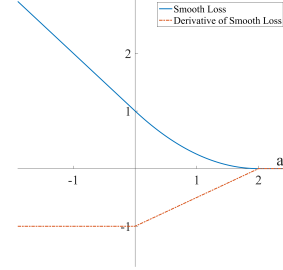


Figure 1: An illustration of the proposed Lipschitz smooth loss function and its derivative.

253 **4.3 The optimization problem of SMILE**

254 Using the classifier defined in Eq. 9, the loss function defined in Eq. 10, and the convex regularization  
 255 terms  $\sum_j \|\mathbf{r}_j^+\|_2^2 + \sum_k \|\mathbf{r}_k^-\|_2^2 + \|\mathbf{L}\|_F^2$ , the following optimization problem is proposed for SMILE:  
 256

$$\min_{\Theta} \frac{1}{n} \sum_{i=1}^n l(y_i h(\mathbf{x}_i; \mathbf{r}^m, \mathbf{L})) + \lambda \left( \sum_{j=1}^{m_+} \|\mathbf{r}_j^+\|_2^2 + \sum_{k=1}^{m_-} \|\mathbf{r}_k^-\|_2^2 + \|\mathbf{L}\|_F^2 \right), \quad (11)$$

257 where  $\Theta = \{\mathbf{r}^m, \mathbf{L}\}$  denotes the set of parameters to be optimized;  $\mathbf{r}^m = \{\mathbf{r}_j^+, \mathbf{r}_k^-; j =$   
 258  $1, \dots, m_+, k = 1, \dots, m_-\}$  denotes the set of representative instances with  $m_+$  instances for  
 259 the positive class and  $m_-$  instances for the negative class; and  $\lambda$  is a trade-off parameter balancing  
 260 the loss term and the regularization term.

261 The objective function is not convex due to the non-convexity of  $h(\mathbf{x}; \mathbf{r}^m, \mathbf{M})$ . We apply the gradient  
 262 descent algorithm to learn the parameters; detailed formulae are given in Appendix D.

263 **4.4 Illustrative results of SMILE**

264 **Experimental settings** We illustrate the effectiveness of SMILE by comparing it with nine widely  
 265 adopted metric learning methods: NCA [16], LMNN [44], ITML [9], R2LML [21], SCML [38],  
 266 RVML [34], GMML [51], DMLMJ [32], and SNC [26]. NCA is implemented by using the drTool-  
 267 box [40]; LMNN and ITML are implemented by using the metric-learn toolbox [10]; and R2LML,  
 268 SCML, RVML, GMML, DMLMJ, and SNC are implemented by using the authors' code.

269 The experiment focuses on binary classification of 12 publicly available datasets from the websites of  
 270 UCI [12] and Delve [36]. Sample size and feature dimension are listed in Table 1 of Appendix E. All  
 271 datasets are pre-processed by firstly subtracting the mean and dividing by the standard deviation, and  
 272 then normalizing the  $L_2$ -norm of each instance to 1.

273 For each dataset, we randomly select 60% of instances to form a training set and the rest are  
 274 used for testing. This process is repeated 10 times and we report the mean accuracy and the  
 275 standard deviation. 10-fold cross-validation is used to select the trade-off parameters in the com-  
 276 pared algorithms, namely the regularization parameter of LMNN (from  $\{0.1, 0.3, \dots, 0.9\}$ ),  $\gamma$  in  
 277 ITML (from  $\{0.25, 0.5, 1, 2, 4\}$ ),  $\lambda$  in RVML (from  $\{10^{-5}, 10^{-4}, \dots, 10\}$ ),  $t$  in GMML (from  
 278  $\{0.1, 0.3, \dots, 0.9\}$ ), and ratio in SNC (from  $\{0.01, 0.02, 0.04, 0.08, 0.16\}$ ). All other parameters are  
 279 set as default. For the proposed SMILE, the parameters are set as follows:  $\mathbf{L}$  is initialized as the  
 280 identity matrix;  $\mathbf{r}^m$  are initialized as the  $k$ -means clustering centers of the positive and negative  
 281 classes (by using MATLAB *kmeans* function with random initial values); the number of representative  
 282 instances for each class is set as 2; the trade-off parameter  $\lambda$  is set as 1; and the learning rate  $\alpha$   
 283 is set as 0.001. The maximum number of iterations is set as 5000 and the final result is based on  
 284 the parameters at time  $t$ , which is the earliest time when the smallest training error is obtained, to  
 285 conform to early stopping as suggested by Theorem 3.



Table 1: Comparison of classification performances. Mean accuracy and standard deviations are reported with the best ones in bold; ‘# of best’ denotes the number of datasets on which the proposed SMILE obtains the highest accuracy.

Dataset	NCA	LMNN	ITML	R2LML	SCML	RVML	GMML	DMLMJ	SNC	SMILE
Australian	80.0±1.6	78.8±2.6	77.2±1.9	84.7±1.3	82.3±1.4	83.0±1.6	84.4±1.0	83.9±1.3	81.8±8.8	<b>86.0±0.7</b>
Cancer	95.4±1.3	96.0±0.7	96.1±1.1	96.7±0.8	96.5±0.5	95.2±1.0	96.5±0.8	96.5±0.5	95.1±1.7	<b>96.8±0.6</b>
Climate	91.5±2.1	91.8±1.3	86.7±1.0	91.7±1.7	91.5±1.5	92.2±1.1	91.3±2.5	92.9±1.9	92.0±1.7	<b>93.5±1.7</b>
Credit	80.6±2.0	82.2±1.4	77.6±2.0	<b>86.1±1.5</b>	83.5±1.2	83.5±1.8	85.9±1.7	84.6±1.4	83.4±3.7	85.6±1.9
German	70.0±2.9	67.9±1.5	67.0±2.1	72.9±1.8	70.9±2.7	71.7±1.8	71.6±1.1	69.3±2.7	70.1±3.3	<b>75.5±1.1</b>
Haberman	67.4±3.3	67.9±3.3	68.0±4.1	71.1±3.4	69.2±2.5	66.7±2.3	71.2±3.4	68.5±3.2	72.0±5.2	<b>72.4±3.3</b>
Heart	75.6±2.0	76.2±3.8	76.9±3.3	82.0±3.8	79.0±3.2	77.7±4.1	81.2±2.7	80.6±2.8	77.0±5.3	<b>84.0±2.2</b>
ILPD	66.8±1.2	67.0±2.1	68.7±2.8	65.9±2.2	68.0±2.9	68.0±2.9	67.1±2.2	68.0±1.6	68.9±2.7	<b>71.3±1.7</b>
Liver	59.8±3.4	61.0±4.8	57.2±4.0	<b>66.8±3.7</b>	61.7±4.6	64.6±3.9	63.8±5.4	60.9±3.8	63.3±5.2	62.8±5.8
Pima	65.9±3.0	68.5±1.6	68.0±2.0	72.3±1.5	71.1±2.6	69.5±1.7	73.0±1.8	71.1±2.3	<b>74.0±2.6</b>	73.2±2.0
Ringnorm	69.3±0.7	65.2±0.7	65.8±0.9	NA	70.9±0.7	72.3±0.6	72.5±0.5	73.9±0.7	71.3±0.6	<b>77.1±0.5</b>
Twonorm	96.7±0.4	95.6±0.5	96.4±0.3	NA	97.3±0.4	97.3±0.3	97.5±0.3	97.7±0.2	97.3±0.2	<b>97.9±0.3</b>
Average	76.6	76.5	75.5	NA	78.5	78.5	78.8	79.7	79.0	<b>81.3</b>
# of best	0	0	0	2	0	0	0	0	1	9

286 **Evaluation on classification performance** As shown in Table 1, with only two representative  
 287 instances learned for each class, the proposed SMILE achieves the best accuracy on 9 out of the 12  
 288 datasets; none of the other methods performs the best on more than 2 datasets. The average accuracy  
 289 of SMILE is also the highest. These results suggest that SMILE, though simple, enjoys competitive  
 290 performance against existing metric learning algorithms, thanks to its theoretical foundation.

291 **Visualization of the concentration behavior** Our theoretical  
 292 finding suggests that randomness of parameters is caused  
 293 by random sampling and will accumulate over iterations. We  
 294 now verify this finding with an empirical study on the dataset  
 295 German. More specifically, we learn parameters  $L, r^m$  from a  
 296 subset of the data, which serves as  $m^{(T)}(z^n)$  in Lemma 2,  
 297 learn parameters from the entire dataset, which serves as  
 298  $\mathbb{E}_{z^n} m^{(T)}(z^n)$ , and quantify their differences via the  $L_2$ -norm.  
 299 The total sample size is 1000 and the subset size is selected  
 300 as  $\{100, 200, \dots, 500\}$ . After randomly sampling the subset  
 301 for 100 times, we calculate the 95th percentile of the norm  
 302 differences and denote this value as  $\epsilon_{95\%}$ .  $\epsilon_{95\%}$  can be interpreted as the minimum radius  $\epsilon$   
 303 of ball  $(\mathbb{E}_{z^n} m^{(T)}(z^n), \epsilon)$  such that the bound (Eq. 6) holds with 95% probability. From Fig. 2, we first  
 304 see that learning from fewer training instances leads to a larger value of  $\epsilon_{95\%}$ , which signifies that  
 305 sampling randomness contributes to the variance of learned parameters. Second, we see that learning  
 306 with more iterations increases  $\epsilon_{95\%}$ , which is also consistent with the theoretical result. Moreover,  
 307 the rate of increase is exponential in the early stage of training and decreases gradually towards zero,  
 308 which implies that parameters are optimized to local minima and will no longer be updated.

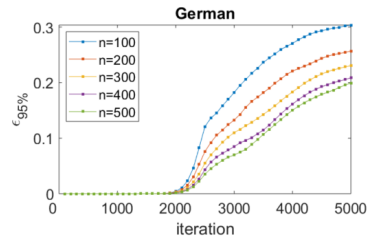


Figure 2: Effect of training iterations and sample size on parameter concentration.

## 309 5 Conclusion

310 This paper presents a new route to the generalization guarantee on classifiers optimized via GD,  
 311 considering the influence of sampling randomness to the concentration property of parameters and  
 312 embracing algorithmic parameters. We propose a new decomposition theorem to obtain the general-  
 313 ization PAC bound, which consequently guarantees the agnostic PAC learnability. We demonstrate  
 314 the necessity of Lipschitz smooth classifiers and loss functions for generalization and theoretically  
 315 justify the benefit of early stopping. Our results are derived based only on the Lipschitz property over  
 316 the parameter space, and hence are applicable to non-convex optimization problems. In addition, we  
 317 propose a new metric learning method as an illustrative example to demonstrate the practicability  
 318 of the appealing theoretical results. In the future, we intend to investigate the link between the  
 319 concentration property and the local convergence behavior, and take it into account to derive tighter  
 320 bounds.

## 321 **Broader Impact**

322 This paper is a theoretical analysis relating to gradient descent and metric learning algorithms, and  
323 hence does not make a direct impact on ethical and societal issues. The findings can be used to  
324 design more effective training strategies or algorithms, and consequently benefit the downstream  
325 applications.

## 326 **References**

- 327 [1] A. Bellet and A. Habrard. Robustness and generalization for metric learning. *Neurocomputing*,  
328 151:259–267, 2015.
- 329 [2] A. Bellet, A. Habrard, and M. Sebban. Similarity learning for provably accurate sparse linear  
330 classification. In *International Conference on International Conference on Machine Learning*,  
331 pages 1491–1498, 2012.
- 332 [3] W. Bian and D. Tao. Constrained empirical risk minimization framework for distance metric  
333 learning. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1194–1205,  
334 2012.
- 335 [4] J. Bohné, Y. Ying, S. Gentric, and M. Pontil. Large margin local metric learning. In *European*  
336 *Conference on Computer Vision*, pages 679–694. Springer, 2014.
- 337 [5] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning*  
338 *Research*, 2(Mar):499–526, 2002.
- 339 [6] Q. Cao, Z.-C. Guo, and Y. Ying. Generalization bounds for metric and similarity learning.  
340 *Machine Learning*, 102(1):115–132, 2016.
- 341 [7] S. Chen, L. Luo, J. Yang, C. Gong, J. Li, and H. Huang. Curvilinear distance metric learning.  
342 In *Advances in Neural Information Processing Systems*, pages 4225–4234, 2019.
- 343 [8] R. Collobert and S. Bengio. Links between perceptrons, mlps and svms. In *International*  
344 *Conference on Machine Learning*, page 23, 2004.
- 345 [9] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In  
346 *International Conference on Machine Learning*, pages 209–216. ACM, 2007.
- 347 [10] W. de Vazelhes, C. Carey, Y. Tang, N. Vauquier, and A. Bellet. metric-learn: Metric Learning  
348 Algorithms in Python. Technical report, arXiv:1908.04710, 2019.
- 349 [11] M. Dong, Y. Wang, X. Yang, and J.-H. Xue. Learning local metrics and influential regions  
350 for classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(6):  
351 1522–1529, 2020.
- 352 [12] D. Dua and C. Graff. UCI machine learning repository, 2017. URL [http://archive.ics.](http://archive.ics.uci.edu/ml)  
353 [uci.edu/ml](http://archive.ics.uci.edu/ml).
- 354 [13] R. M. Dudley. *Uniform central limit theorems*, volume 142. Cambridge University Press, 2014.
- 355 [14] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions  
356 for shape-based image retrieval and classification. In *International Conference on Computer*  
357 *Vision*, pages 1–8. IEEE, 2007.
- 358 [15] L. Gautheron, A. Habrard, E. Morvant, and M. Sebban. Metric learning from imbalanced data  
359 with generalization guarantees. *Pattern Recognition Letters*, 133:298–304, 2020.
- 360 [16] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov. Neighbourhood components  
361 analysis. In *Advances in Neural Information Processing Systems*, pages 513–520, 2005.
- 362 [17] L. A. Gottlieb, A. Kontorovich, and R. Krauthgamer. Efficient classification for metric data.  
363 *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014.

- 364 [18] Z.-C. Guo and Y. Ying. Guaranteed classification via regularized similarity learning. *Neural*  
365 *Computation*, 26(3):497–522, 2014.
- 366 [19] C. Gupta, A. S. Suggala, A. Goyal, H. V. Simhadri, B. Paranjape, A. Kumar, S. Goyal, R. Udupa,  
367 M. Varma, and P. Jain. ProtoNN: Compressed and accurate kNN for resource-scarce devices.  
368 In *International Conference on Machine Learning*, pages 1331–1340, 2017.
- 369 [20] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient  
370 descent. *ICML*, 2016.
- 371 [21] Y. Huang, C. Li, M. Georgiopoulos, and G. C. Anagnostopoulos. Reduced-rank local distance  
372 metric learning. In *Joint European Conference on Machine Learning and Knowledge Discovery*  
373 *in Databases*, pages 224–239. Springer, 2013.
- 374 [22] Z. Huang, R. Wang, S. Shan, and X. Chen. Projection metric learning on grassmann manifold  
375 with application to video based face recognition. In *IEEE Conference on Computer Vision and*  
376 *Pattern Recognition*, pages 140–149, 2015.
- 377 [23] P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*,  
378 pages 73–101, 1964.
- 379 [24] Z. Huo, F. Nie, and H. Huang. Robust and effective metric learning using capped trace  
380 norm: Metric learning via capped trace norm. In *ACM SIGKDD International Conference on*  
381 *Knowledge Discovery and Data Mining*, pages 1605–1614, 2016.
- 382 [25] R. Jin, S. Wang, and Y. Zhou. Regularized distance metric learning: Theory and algorithm. In  
383 *Advances in Neural Information Processing Systems*, pages 862–870, 2009.
- 384 [26] M. Kusner, S. Tyree, K. Q. Weinberger, and K. Agrawal. Stochastic neighbor compression. In  
385 *International Conference on Machine Learning*, pages 622–630, 2014.
- 386 [27] D. Lim, G. Lanckriet, and B. McFee. Robust structural metric learning. In *International*  
387 *Conference on Machine Learning*, pages 615–623, 2013.
- 388 [28] U. v. Luxburg and O. Bousquet. Distance-based classification with lipschitz functions. *Journal*  
389 *of Machine Learning Research*, 5(Jun):669–695, 2004.
- 390 [29] B. Mason, L. Jain, and R. Nowak. Learning low-dimensional metrics. In *Advances in Neural*  
391 *Information Processing Systems*, pages 4139–4147, 2017.
- 392 [30] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh. No fuss distance metric  
393 learning using proxies. In *IEEE International Conference on Computer Vision*, pages 360–368,  
394 2017.
- 395 [31] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep  
396 learning. In *Advances in Neural Information Processing Systems*, pages 5949–5958, 2017.
- 397 [32] B. Nguyen, C. Morell, and B. De Baets. Supervised distance metric learning through maximiza-  
398 tion of the jeffrey divergence. *Pattern Recognition*, 64:215–225, 2017.
- 399 [33] Y.-K. Noh, M. Sugiyama, K.-E. Kim, F. Park, and D. D. Lee. Generative local metric learning  
400 for kernel regression. In *Advances in Neural Information Processing Systems*, pages 2452–2462,  
401 2017.
- 402 [34] M. Perrot and A. Habrard. Regressive virtual metric learning. In *Advances in Neural Information*  
403 *Processing Systems*, pages 1810–1818, 2015.
- 404 [35] Q. Qian, J. Tang, H. Li, S. Zhu, and R. Jin. Large-scale distance metric learning with uncertainty.  
405 In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8542–8550, 2018.
- 406 [36] C. Rasmussen, R. Neal, G. Hinton, D. Van Camp, M. Revow, Z. Ghahramani, R. Kustra,  
407 and R. Tibshirani. Delve data for evaluating learning in valid experiments. URL [http:](http://www.cs.toronto.edu/~delve)  
408 [//www.cs.toronto.edu/~delve](http://www.cs.toronto.edu/~delve).

- 409 [37] K. Roth, T. Milbich, S. Sinha, P. Gupta, B. Ommer, and J. P. Cohen. Revisiting training  
410 strategies and generalization performance in deep metric learning. In *International Conference*  
411 *on Machine Learning*, 2020.
- 412 [38] Y. Shi, A. Bellet, and F. Sha. Sparse compositional metric learning. In *AAAI Conference on*  
413 *Artificial Intelligence*, pages 2078–2084, 2014.
- 414 [39] M. Taheri, Z. Moslehi, A. Mirzaei, and M. Safayani. A self-adaptive local metric learning  
415 method for classification. *Pattern Recognition*, 96:106994, 2019.
- 416 [40] L. van der Maaten. Matlab toolbox for dimensionality reduction. URL [https://](https://lvdmaaten.github.io/drtoolbox/)  
417 [lvdmaaten.github.io/drtoolbox/](https://lvdmaaten.github.io/drtoolbox/).
- 418 [41] N. Verma and K. Branson. Sample complexity of learning Mahalanobis distance metrics. In  
419 *Advances in Neural Information Processing Systems*, pages 2584–2592, 2015.
- 420 [42] J. Wang, A. Kalousis, and A. Woznica. Parametric local metric learning for nearest neighbor  
421 classification. In *Advances in Neural Information Processing Systems*, pages 1601–1609, 2012.
- 422 [43] N. Weaver. *Lipschitz Algebras*. World Scientific, 1999.
- 423 [44] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor  
424 classification. *The Journal of Machine Learning Research*, 10:207–244, 2009.
- 425 [45] P. Xie, W. Wu, Y. Zhu, and E. Xing. Orthogonality-promoting distance metric learning: Convex  
426 relaxation and theoretical analysis. In *International Conference on Machine Learning*, pages  
427 5403–5412, 2018.
- 428 [46] E. P. Xing, M. I. Jordan, S. Russell, and A. Y. Ng. Distance metric learning with application to  
429 clustering with side-information. In *Advances in Neural Information Processing Systems*, pages  
430 505–512, 2002.
- 431 [47] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning.  
432 *Constructive Approximation*, 26(2):289–315, 2007.
- 433 [48] H.-J. Ye, D.-C. Zhan, and Y. Jiang. Fast generalization rates for distance metric learning.  
434 *Machine Learning*, 108(2):267–295, 2019.
- 435 [49] H.-J. Ye, D.-C. Zhan, N. Li, and Y. Jiang. Learning multiple local metrics: Global consideration  
436 helps. *IEEE transactions on pattern analysis and machine intelligence*, 2019.
- 437 [50] Y. Ying, K. Huang, and C. Campbell. Sparse metric learning via smooth optimization. In  
438 *Advances in Neural Information Processing Systems*, pages 2214–2222, 2009.
- 439 [51] P. Zadeh, R. Hosseini, and S. Sra. Geometric mean metric learning. In *International Conference*  
440 *on Machine Learning*, pages 2464–2471, 2016.
- 441 [52] K. Zhong, R. Guo, S. Kumar, B. Yan, D. Simcha, and I. Dhillon. Fast classification with binary  
442 prototypes. In *Artificial Intelligence and Statistics*, pages 1255–1263, 2017.
- 443 [53] P. Zhu, H. Cheng, Q. Hu, Q. Wang, and C. Zhang. Towards generalized and efficient metric  
444 learning on riemannian manifold. In *International Joint Conference on Artificial Intelligence*,  
445 pages 3235–3241, 2018.

---

# Generalization Bound of Gradient Descent for Non-Convex Metric Learning

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1        Sec. A lists properties of Lipschitz functions. Sec. B includes established definitions  
2        and theorems that will be used in the proof, including definitions of PAC learnability  
3        and agnostic PAC learnability, the McDiarmid’s Inequality. Sec. C provides proofs  
4        of theorems and lemmas. Sec. D shows that the classifier of SMILE is smooth and  
5        gives the updating equations of the gradient descent algorithm. Sec. E lists the  
6        information about the datasets.

## 7    A    Properties of Lipschitz functions

8        The Lipschitz constant of differentiable functions can be obtained from their gradients; this follows  
9        from the mean value theorem as shown below.

10      **Theorem A.1.** [3] Let  $\mathcal{U} \in \mathbb{R}^n$  be open,  $h : \mathcal{U} \rightarrow \mathbb{R}$  be differentiable and the line segment  
11       $[\mathbf{u}_1, \mathbf{u}_2] \in \mathcal{U}$ , where  $[\mathbf{u}_1, \mathbf{u}_2] = \{\mathbf{v} \mid \mathbf{v} = \mathbf{u}_1 + t(\mathbf{u}_2 - \mathbf{u}_1), t \in [0, 1]\}$  joins  $\mathbf{u}_1$  to  $\mathbf{u}_2$ . Based on the  
12      *Mean Value Theorem*, there exists a  $\mathbf{u} \in [\mathbf{u}_1, \mathbf{u}_2]$

$$f(\mathbf{u}_2) - f(\mathbf{u}_1) = f'(\mathbf{u})^T(\mathbf{u}_2 - \mathbf{u}_1).$$

13      **Corollary A.1.** Let  $\mathcal{U} \in \mathbb{R}^n$  be open and convex,  $h : \mathcal{U} \rightarrow \mathbb{R}$  be differentiable inside  $\mathcal{U}$ , then the  
14      following inequality holds:

$$\text{lip}(h \leftarrow \mathbf{u}) = \max_{\mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U}, \mathbf{u}_1 \neq \mathbf{u}_2} \frac{|h(\mathbf{u}_2) - h(\mathbf{u}_1)|}{\|\mathbf{u}_2 - \mathbf{u}_1\|} \leq \max_{\mathbf{u} \in \mathcal{U}} \|h'(\mathbf{u})\|.$$

15      *Proof.* Since  $\mathcal{U}$  is convex,  $\forall \mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U}, \mathbf{u}_1 \neq \mathbf{u}_2$ , the line segment  $[\mathbf{u}_1, \mathbf{u}_2] = \{\mathbf{v} \mid \mathbf{v} = \mathbf{u}_1 +$   
16       $t(\mathbf{u}_2 - \mathbf{u}_1), t \in [0, 1]\} \in \mathcal{U}$ .

$$|h(\mathbf{u}_2) - h(\mathbf{u}_1)| =_{(a)} |h'(\mathbf{u})^T(\mathbf{u}_2 - \mathbf{u}_1)| \leq_{(b)} \|h'(\mathbf{u})\| \|\mathbf{u}_2 - \mathbf{u}_1\| \leq_{(c)} \max_{\mathbf{u} \in \mathcal{U}} \|h'(\mathbf{u})\| \|\mathbf{u}_2 - \mathbf{u}_1\|,$$

17      where equality (a) is due to Theorem A.1; inequality (b) is due to the Cauchy-Schwarz inequality;  
18      inequality (c) is due to  $\|h'(\mathbf{u})\| \leq \max_{\mathbf{u} \in \mathcal{U}} \|h'(\mathbf{u})\|$ .  $\square$

19      Sophisticated Lipschitz functions can be constructed from the basic ones using the following lemma.

20      **Lemma A.1.** [4, 10] Let  $\text{lip}(h_1 \leftarrow \mathbf{u}) \leq L_1$ ,  $\text{lip}(h_2 \leftarrow \mathbf{u}) \leq L_2$  and  $\text{lip}(h_2 \circ h_1 \leftarrow h_1) \leq L_3$ ,  
21      where  $\circ$  denotes the composition of functions. Then

- 22      (a)  $\text{lip}(ah_1 \leftarrow \mathbf{u}) \leq |a|L_1$ , where  $a$  is a constant;  
23      (b)  $\text{lip}(h_1 + h_2 \leftarrow \mathbf{u}) \leq L_1 + L_2$ ,  $\text{lip}(h_1 - h_2 \leftarrow \mathbf{u}) \leq L_1 + L_2$ ;  
24      (c)  $\text{lip}(\min(h_1, h_2) \leftarrow \mathbf{u}) \leq \max\{L_1, L_2\}$ ,  $\text{lip}(\max(h_1, h_2) \leftarrow \mathbf{u}) \leq \max\{L_1, L_2\}$ , where  
25       $\min(h_1, h_2)$  or  $\max(h_1, h_2)$  denote the pointwise minimum or maximum of functions  $h_1$  and  $h_2$ ;  
26      (d)  $\text{lip}(h_2 \circ h_1 \leftarrow \mathbf{u}) \leq L_1 L_3$ .

27      This lemma illustrates that after the operations of multiplication by constant, addition, subtraction,  
28      minimization, maximization and function composition, the functions are still Lipschitz continuous.

## 29 B Preliminaries

30 **Definition B.1.** [6, 8] A hypothesis class  $\mathcal{H}$  is *Probably Approximately Correct (PAC) learnable* if  
 31 there exist a function  $n_{\mathcal{H}}^L : (0, 1)^2 \rightarrow \mathbb{N}$  and a learning algorithm with the following property: For  
 32 every  $\epsilon, \delta \in (0, 1)$ , for every distribution  $\mathcal{D}_{\mathcal{X}}$  over  $\mathcal{X}$ , and for every target function  $g \in \mathcal{G}$ , if there  
 33 exists an  $h^* \in \mathcal{H}$  which returns the same classification result as  $g$ , then when running the learning  
 34 algorithm on  $n \geq n_{\mathcal{H}}^L(\epsilon, \delta)$  independent and identically distributed (i.i.d.) instances generated by  
 35  $\mathcal{D}_{\mathcal{X}}$  and labeled by  $g$ , the algorithm returns a hypothesis  $\hat{h}$  such that, with probability at least  $1 - \delta$ ,  
 36  $R(\hat{h}) \leq \epsilon$ ; this can be equivalently written as

$$\mathbb{P}_{\mathbf{x}^n} [R(\hat{h}) \leq \epsilon] \geq 1 - \delta,$$

37 or

$$\mathbb{P}_{\mathbf{x}^n} \left[ \mathbb{E}_{\mathbf{x}'} [l(\hat{h}(\mathbf{x}'), g(\mathbf{x}'))] \leq \epsilon \right] \geq 1 - \delta,$$

38 where the probability is taken over  $\mathbf{x}_n$  and  $\hat{h}$  is a random variable related to  $\mathbf{x}_n$ .

39 **Definition B.2.** [6, 2] A hypothesis class  $\mathcal{H}$  is *agnostic PAC learnable* or has *agnostic PAC learnability*  
 40 if there exist a function  $n_{\mathcal{H}}^{AL} : (0, 1)^2 \rightarrow \mathbb{N}$  and a learning algorithm with the following property:  
 41 For every  $\epsilon, \delta \in (0, 1)$  and for every distribution  $\mathcal{D}_{\mathcal{Z}}$  over  $\mathcal{Z}$ , when running the learning algorithm on  
 42  $n \geq n_{\mathcal{H}}^{AL}(\epsilon, \delta)$  i.i.d. instances generated by  $\mathcal{D}_{\mathcal{Z}}$ , the algorithm returns a hypothesis  $\hat{h}$  such that, with  
 43 probability at least  $1 - \delta$ ,  $R(\hat{h}) - \min_{h \in \mathcal{H}} R(h) \leq \epsilon$ ; this can be equivalently written as

$$\mathbb{P}_{\mathbf{z}^n} \left[ R(\hat{h}) - \min_{h \in \mathcal{H}} R(h) \leq \epsilon \right] \geq 1 - \delta,$$

44 or

$$\mathbb{P}_{\mathbf{z}^n} \left[ \mathbb{E}_{\mathbf{z}'} [l(\hat{h}(\mathbf{x}'), y)] - \min_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{z}'} [l(h(\mathbf{x}'), y)] \leq \epsilon \right] \geq 1 - \delta.$$

45 where the probability is taken over  $\mathbf{z}_n$  and  $\hat{h}$  is a random variable related to  $\mathbf{z}_n$ .

46 **Lemma B.1.** [5, (McDiarmid's Inequality)] Let  $\mathbf{z}^n = \{z_1, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_n\}$  be  $n$  inde-  
 47 pendent samples. Let  $\mathbf{z}^{n,i} = \{z_1, \dots, z_{i-1}, z'_i, z_{i+1}, \dots, z_n\}$ , where the replacement example  $z'_i$  is  
 48 assumed to be drawn from the same distribution of  $z_i$  and is independent from  $\mathbf{z}^n$ . Furthermore, let  
 49  $m : \mathcal{Z}^n \rightarrow \mathbb{R}$  be a function of  $z_1, \dots, z_n$  that satisfies  $\forall i, \forall \mathbf{z}^n, \forall \mathbf{z}^{n,i}$

$$|m(\mathbf{z}^n) - m(\mathbf{z}^{n,i})| \leq c_i, \quad (1)$$

50 for some constant  $c_i$ . Then for all  $\epsilon > 0$ ,

$$\mathbb{P}_{\mathbf{z}^n} [m(\mathbf{z}^n) - \mathbb{E}_{\mathbf{z}^n} [m(\mathbf{z}^n)] \geq \epsilon] \leq \exp \left( \frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2} \right),$$

51

$$\mathbb{P}_{\mathbf{z}^n} [\mathbb{E}_{\mathbf{z}^n} [m(\mathbf{z}^n)] - m(\mathbf{z}^n) \geq \epsilon] \leq \exp \left( \frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2} \right);$$

52 that is,

$$\mathbb{P}_{\mathbf{z}^n} [ |m(\mathbf{z}^n) - \mathbb{E}_{\mathbf{z}^n} [m(\mathbf{z}^n)]| \geq \epsilon ] \leq 2 \exp \left( \frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2} \right). \quad (2)$$

## 53 C Proofs of theorems and lemmas

### 54 C.1 Proof of Lemma 1

55 *Proof.* Let  $E_1$  be the set of events of  $R(\hat{h}) - R_n(\mathbf{z}^n, \hat{h}) \leq \epsilon$  and  $E_2$  be the set of events of  
 56  $\max_{h \in \mathcal{H}} (R(h) - R_n(\mathbf{z}^n, h)) \leq \epsilon$ . The probabilities of these two events are given as follows:

$$\begin{aligned} \mathbb{P}_{\mathbf{z}^n}(E_1) &= \int (p(\mathbf{z}^n) \mathbb{1}[E_1]) d\mathbf{z}^n \\ \mathbb{P}_{\mathbf{z}^n}(E_2) &= \int (p(\mathbf{z}^n) \mathbb{1}[E_2]) d\mathbf{z}^n. \end{aligned}$$

57 1. At the points  $\mathbf{z}^n$  where  $\mathbb{1}[E_2] = 1$ , we have  $\mathbb{1}[E_1] = 1$  and thus

$$p(\mathbf{z}^n)\mathbb{1}[E_1] = p(\mathbf{z}^n)\mathbb{1}[E_2].$$

58 2. At the points  $\mathbf{z}^n$  where  $\mathbb{1}[E_2] = 0$ , we have

$$p(\mathbf{z}^n)\mathbb{1}[E_1] \geq 0 = p(\mathbf{z}^n)\mathbb{1}[E_2].$$

59 Therefore, integrating over all possible points  $\mathbf{z}^n$ , we have  $(\int p(\mathbf{z}^n)\mathbb{1}[E_1]) \geq (\int p(\mathbf{z}^n)\mathbb{1}[E_2])$ . That  
60 is,  $\mathbb{P}_{\mathbf{z}^n}(E_1) \geq \mathbb{P}_{\mathbf{z}^n}(E_2)$ .  $\square$

## 61 C.2 Proof of Theorem 1

62 After proving Proposition C.1, Theorem 1 is proved.

63 **Proposition C.1.** Suppose the range of the risk function  $r(\mathbf{z}, h)$  is  $[0, C_r]$ , then

$$\mathbb{P}_{\mathbf{z}^n} \left[ \min_{h \in \mathcal{H}} R_n(\mathbf{z}^n, h) - E_{\mathbf{z}^n} \left[ \min_{h \in \mathcal{H}} R_n(\mathbf{z}^n, h) \right] \geq \epsilon \right] \leq \exp \left( \frac{-2n\epsilon^2}{C_r^2} \right).$$

64 *Proof.* Given  $\mathbf{z}^n$  and a fixed hypothesis class of  $\mathcal{H}$ , the value of  $a(\mathbf{z}^n) = \min_{h \in \mathcal{H}} R_n(\mathbf{z}^n, h)$   
65 is fixed and the mapping  $a : \mathcal{Z}^n \rightarrow \mathbb{R}$  is a function. Therefore, the McDiarmid's inequality  
66 (Lemma B.1) can be applied as long as the bounded difference condition (Eq. 1) holds. We show that  
67  $|\min_{h \in \mathcal{H}} R_n(\mathbf{z}^n, h) - \min_{h \in \mathcal{H}} R_n(\mathbf{z}^{n,i}, h)|$  is bounded as follows:

$$\begin{aligned} & \min_{h \in \mathcal{H}} R_n(\mathbf{z}^{n,i}, h) \\ &= \min_{h \in \mathcal{H}} \left( R_n(\mathbf{z}^n, h) - \frac{r(z_i, h)}{n} + \frac{r(z'_i, h)}{n} \right) \\ &\leq \min_{h \in \mathcal{H}} \left( R_n(\mathbf{z}^n, h) - 0 + \frac{C_r}{n} \right) \\ &= \min_{h \in \mathcal{H}} R_n(\mathbf{z}^n, h) + \frac{C_r}{n}. \end{aligned}$$

68 Similarly,

$$\min_{h \in \mathcal{H}} R_n(\mathbf{z}^n, h) \leq \min_{h \in \mathcal{H}} R_n(\mathbf{z}^{n,i}, h) + \frac{C_r}{n}.$$

69 Therefore

$$\left| \min_{h \in \mathcal{H}} R_n(\mathbf{z}^n, h) - \min_{h \in \mathcal{H}} R_n(\mathbf{z}^{n,i}, h) \right| \leq \frac{C_r}{n}.$$

70 The result is obtained by substituting  $c_i = \frac{C_r}{n}$  into Lemma B.1.  $\square$

71 Theorem 1 is proved as follows.

72 *Proof.* Let  $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} R_n(\mathbf{z}^n, h)$ , we have

$$R_n(\mathbf{z}^n, \hat{h}) = \min_{h \in \mathcal{H}} R_n(\mathbf{z}^n, h).$$

73 Suppose

$$\mathbb{P}_{\mathbf{z}^n} [R(\hat{h}) - R_n(\mathbf{z}^n, \hat{h}) \leq \epsilon/2] \geq 1 - \delta/2,$$

74

$$\mathbb{P}_{\mathbf{z}^n} [R_n(\mathbf{z}^n, \hat{h}) - E_{\mathbf{z}^n} [R_n(\mathbf{z}^n, \hat{h})] \leq \epsilon/2] \geq 1 - \delta/2.$$

75 Let  $E_1 = \{\mathbf{z}^n | R(\hat{h}) - R_n(\mathbf{z}^n, \hat{h}) \leq \epsilon/2\}$  and  $E_2 = \{\mathbf{z}^n | R_n(\mathbf{z}^n, \hat{h}) - E_{\mathbf{z}^n}[R_n(\mathbf{z}^n, \hat{h})] \leq \epsilon/2\}$ .  
 76  $\forall \mathbf{z}^n \in E_1 \cap E_2$ , we have

$$\begin{aligned}
 & R(\hat{h}) \\
 (a) \quad & \leq R_n(\mathbf{z}^n, \hat{h}) + \frac{\epsilon}{2} \\
 (b) \quad & \leq \mathbb{E}_{\mathbf{z}^n}[R_n(\mathbf{z}^n, \hat{h})] + \epsilon \\
 (c) \quad & = \mathbb{E}_{\mathbf{z}^n} \min_{h \in \mathcal{H}} \frac{\sum_{i=1}^n r(\mathbf{z}_i, h)}{n} + \epsilon \\
 (d) \quad & \leq \min_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{z}^n} \frac{\sum_{i=1}^n r(\mathbf{z}_i, h)}{n} + \epsilon \\
 (e) \quad & = \min_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{z}^n} r(\mathbf{z}, h) + \epsilon \\
 (f) \quad & = \min_{h \in \mathcal{H}} R(h) + \epsilon,
 \end{aligned}$$

77 where inequality (a) is due to  $R(\hat{h}) - R_n(\mathbf{z}^n, \hat{h}) \leq \epsilon/2$ ; inequality (b) is due to  $R_n(\mathbf{z}^n, \hat{h}) -$   
 78  $E_{\mathbf{z}^n}[R_n(\mathbf{z}^n, \hat{h})] \leq \epsilon/2$ ; equality (c) is due to the definitions of  $R_n(\mathbf{z}^n, h)$  and  $\hat{h}$ ; inequality (d) is  
 79 due to change the order of  $E_{\mathbf{z}^n}$  and  $\min_{h \in \mathcal{H}}$ ; equality (e) is due to the identical assumption of  $\mathbf{z}^n$ ;  
 80 equality (f) is due to the definition of  $R(h)$ .

81 Therefore

$$\begin{aligned}
 & \mathbb{P}_{\mathbf{z}^n} \left[ R(\hat{h}) \leq \min_{h \in \mathcal{H}} R(h) + \epsilon \right] \\
 (a) \quad & \geq \mathbb{P}_{\mathbf{z}^n} [E_1 \cap E_2] \\
 (b) \quad & \geq 1 - \delta/2 - \delta/2,
 \end{aligned}$$

82 where inequality (a) is due to the relationship between  $E_1 \cap E_2$  and  $R(\hat{h}) \leq \min_{h \in \mathcal{H}} R(h) + \epsilon$ ;  
 83 inequality (b) is due to the probability of union of sets.

84 Based on Proposition C.1, in order to guarantee  $\mathbb{P}_{\mathbf{z}^n} [R_n(\mathbf{z}^n, \hat{h}) - E_{\mathbf{z}^n}[R_n(\mathbf{z}^n, \hat{h})] \leq \epsilon/2] \geq 1 - \delta/2$ ,  
 85  $\frac{2C_r^2}{\epsilon^2} \ln \frac{4}{\delta}$  instances are required. Meanwhile, based on the definition of generalization PAC bound  
 86 (Definition 4 of the main text), in order to guarantee  $\mathbb{P}_{\mathbf{z}^n} [R(\hat{h}) - R_n(\mathbf{z}^n, \hat{h}) \leq \epsilon/2] \geq 1 - \delta/2$ ,  
 87  $m_{\mathcal{H}}^G(\epsilon/2, \delta/2)$  instances are required. Therefore, with more than  $\max(m_{\mathcal{H}}^G(\epsilon/2, \delta/2), \frac{2C_r^2}{\epsilon^2} \ln \frac{4}{\delta})$   
 88 instances,  $\mathbb{P}_{\mathbf{z}^n} [R(\hat{h}) \leq \min_{h \in \mathcal{H}} R(h) + \epsilon] \geq 1 - \delta$  is satisfied. Based on the definition of the  
 89 agnostic PAC learnability (Definition B.2), the hypothesis class is agnostic PAC learnable and the  
 90 agnostic PAC learner for  $\mathcal{H}$  is  $\text{ERM}_{\mathcal{H}}$ .  $\square$

### 91 C.3 Proof of Theorem 2

92 *Proof.* Let  $E_1$  denote the set of events  $R(\hat{\mathbf{w}}) - R_n(\mathbf{z}^n, \hat{\mathbf{w}}) \leq \epsilon$ ,  $E_2$  denote the set of events  $\mathbf{w} \in \hat{\mathcal{W}}$ ,  
 93 and  $E_3$  denote the set of events  $\max_{\mathbf{w} \in \hat{\mathcal{W}}} [R(\mathbf{w}) - R_n(\mathbf{z}^n, \mathbf{w})] \leq \epsilon$ .

$$\begin{aligned}
 & \mathbb{P}_{\mathbf{z}^n} [\neg E_1] \\
 & = \mathbb{P}_{\mathbf{z}^n} [\neg E_1, E_2] + \mathbb{P}_{\mathbf{z}^n} [\neg E_1, \neg E_2] \\
 (a) \quad & \leq \mathbb{P}_{\mathbf{z}^n} [\neg E_1, E_2] + \delta_1 \\
 (b) \quad & \leq \mathbb{P}_{\mathbf{z}^n} [\neg E_3] + \delta_1 \\
 & = \delta_2 + \delta_1;
 \end{aligned}$$

94 where inequality (a) is due to  $\mathbb{P}_{\mathbf{z}^n} [\neg E_1, \neg E_2] \leq \mathbb{P}_{\mathbf{z}^n} [\neg E_2] = 1 - \mathbb{P}_{\mathbf{z}^n} [E_2] \leq \delta_1$ ; inequality (b) is  
 95 based on the relationship between  $\mathbb{1}[E_2] \mathbb{1}[\neg E_1]$  and  $\mathbb{1}[E_3]$ . At the points  $\mathbf{z}^n$  that satisfy  $\mathbf{m}(\mathbf{z}^n) =$   
 96  $\hat{\mathbf{w}} \in \hat{\mathcal{W}}$ ,  $\mathbb{1}[\neg E_1] = 1 \Rightarrow \mathbb{1}[\neg E_3] = 1$ , thus  $\mathbb{1}[E_2] \mathbb{1}[\neg E_1] \leq \mathbb{1}[\neg E_3]$  and  $\mathbb{P}_{\mathbf{z}^n} [\neg E_1, E_2] \leq$   
 97  $\mathbb{P}_{\mathbf{z}^n} [\neg E_3]$ .  $\square$

### 98 C.4 Proof of Lemma 2

99 *Proof.* To show that  $\mathbf{m}_{[q]}^{(T)}(\mathbf{z}^n; \mathbf{s})$  is concentrated around its expectation, we make use of the Mc-  
 100 Diarmid's Inequality (Lemma B.1). First, we note that  $\mathbf{m}_{[q]}^{(T)}(\mathbf{z}^n; \mathbf{s}) : \mathcal{Z}^n \rightarrow \mathbb{R}$  is function map-



101 ping from random variables to a real value, and  $\mathbf{z}^n$  satisfies the independent assumption. Second,  
 102 we show that  $\|\mathbf{m}_{[q]}^{(T)}(\mathbf{z}^n; \mathbf{s}) - \mathbf{m}_{[q]}^{(T)}(\mathbf{z}^{n,i}; \mathbf{s})\|$  is bounded.  $\mathbf{m}^{(T)}(\mathbf{z}^n; \mathbf{s})$  and  $\mathbf{m}_{[q]}^{(T)}(\mathbf{z}^n; \mathbf{s})$  are tem-  
 103 porarily simplified to  $\mathbf{m}^{(T)}(\mathbf{z}^n)$  and  $\mathbf{m}_{[q]}^{(T)}(\mathbf{z}^n)$ , respectively.  $\forall \mathbf{s}, \forall q, \|\mathbf{m}_{[q]}^{(T)}(\mathbf{z}^n) - \mathbf{m}_{[q]}^{(T)}(\mathbf{z}^{n,i})\| \leq$   
 104  $\|\mathbf{m}^{(T)}(\mathbf{z}^n) - \mathbf{m}^{(T)}(\mathbf{z}^{n,i})\|$ , where  $\|\cdot\|$  denotes the vector  $L_2$ -norm<sup>1</sup>. We will now discuss the bound  
 105 of  $\|\mathbf{m}^{(T)}(\mathbf{z}^n) - \mathbf{m}^{(T)}(\mathbf{z}^{n,i})\|$ .

106 (1) Decompose  $\mathbf{m}^{(t)}(\mathbf{z}^n)$ . To understand the influence of  $\mathbf{z}_i$ , the updating equation of  $\mathbf{m}^{(t)}(\mathbf{z}^n)$  is  
 107 divided into two parts:

$$\mathbf{m}^{(t)}(\mathbf{z}^n) = \left( \mathbf{m}^{(t-1)}(\mathbf{z}^n) - \sum_{j \in [n]/i} \frac{\alpha^{(t)}}{n} \frac{\partial r(\mathbf{z}_j, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{m}^{(t-1)}(\mathbf{z}^n)} \right) - \frac{\alpha^{(t)}}{n} \frac{\partial r(\mathbf{z}_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{m}^{(t-1)}(\mathbf{z}^n)}.$$

108 Representing the above updating process via the function  $G$  gives:

$$\mathbf{m}^{(t)}(\mathbf{z}^n) = G(\mathbf{m}^{(t-1)}(\mathbf{z}^n)) - \frac{\alpha^{(t)}}{n} \frac{\partial r(\mathbf{z}_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{m}^{(t-1)}(\mathbf{z}^n)}.$$

109 For both  $\mathbf{z}^n$  and  $\mathbf{z}^{n,i}$ ,  $G(\mathbf{m}^{(t-1)}(\mathbf{z}^n)) = G(\mathbf{m}^{(t-1)}(\mathbf{z}^{n,i}))$  because the training instances considered  
 110 in  $G$  are the same. Then

$$\begin{aligned} & \|\mathbf{m}^{(t)}(\mathbf{z}^n) - \mathbf{m}^{(t)}(\mathbf{z}^{n,i})\| \\ &= \left\| G(\mathbf{m}^{(t-1)}(\mathbf{z}^n)) - \frac{\alpha^{(t)}}{n} \frac{\partial r(\mathbf{z}_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{m}^{(t-1)}(\mathbf{z}^n)} - \right. \\ & \quad \left. G(\mathbf{m}^{(t-1)}(\mathbf{z}^{n,i})) + \frac{\alpha^{(t)}}{n} \frac{\partial r(\mathbf{z}'_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{m}^{(t-1)}(\mathbf{z}^{n,i})} \right\| \\ &\leq \left\| \frac{\alpha^{(t)}}{n} \frac{\partial r(\mathbf{z}_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{m}^{(t-1)}(\mathbf{z}^n)} - \frac{\alpha^{(t)}}{n} \frac{\partial r(\mathbf{z}'_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{m}^{(t-1)}(\mathbf{z}^{n,i})} \right\| \text{ (Term 1)} + \\ & \quad \|G(\mathbf{m}^{(t-1)}(\mathbf{z}^n)) - G(\mathbf{m}^{(t-1)}(\mathbf{z}^{n,i}))\| \text{ (Term 2)}. \end{aligned}$$

111 Term 1 and Term 2 in the inequality can be bounded by using the Lipschitz constant of a function  $r$   
 112 with respect to  $\mathbf{w}$  and the Lipschitz constant of  $G$  with respect to  $\mathbf{w}$ , respectively.

113 (2) Bound Term 1. Recall that the Lipschitz constant is defined as:

$$\text{lip}(r \leftarrow \mathbf{w}) = \max_{\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}, \mathbf{w}_1 \neq \mathbf{w}_2, \mathbf{z} \in \mathcal{Z}} \frac{|r(\mathbf{z}; \mathbf{w}_1) - r(\mathbf{z}; \mathbf{w}_2)|}{\|\mathbf{w}_1 - \mathbf{w}_2\|}.$$

114 Term 1 is bounded as follows:

$$\begin{aligned} & \left\| \frac{\alpha^{(t)}}{n} \frac{\partial r(\mathbf{z}_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{m}^{(t-1)}(\mathbf{z}^n)} - \frac{\alpha^{(t)}}{n} \frac{\partial r(\mathbf{z}'_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{m}^{(t-1)}(\mathbf{z}^{n,i})} \right\| \\ &\leq \left\| \frac{\alpha^{(t)}}{n} \frac{\partial r(\mathbf{z}_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{m}^{(t-1)}(\mathbf{z}^n)} \right\| + \left\| \frac{\alpha^{(t)}}{n} \frac{\partial r(\mathbf{z}'_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{m}^{(t-1)}(\mathbf{z}^{n,i})} \right\| \\ &\leq \frac{2\alpha^{(t)}}{n} \text{lip}(r \leftarrow \mathbf{w}). \end{aligned}$$

115 (3) Bound Term 2. Let  $\eta = \text{lip}(G \leftarrow \mathbf{w})$ .

$$\|G(\mathbf{m}^{(t-1)}(\mathbf{z}^n)) - G(\mathbf{m}^{(t-1)}(\mathbf{z}^{n,i}))\| \leq \eta \|\mathbf{m}^{(t-1)}(\mathbf{z}^n) - \mathbf{m}^{(t-1)}(\mathbf{z}^{n,i})\|$$

116 (4) Bound  $\|\mathbf{m}^{(T)}(\mathbf{z}^n) - \mathbf{m}^{(T)}(\mathbf{z}^{n,i})\|$

117  $t = 1$

$$\begin{aligned} & \|\mathbf{m}^{(1)}(\mathbf{z}^n) - \mathbf{m}^{(1)}(\mathbf{z}^{n,i})\| \\ &\leq \left\| \frac{\alpha^{(1)}}{n} \frac{\partial r(\mathbf{z}_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}^0} - \frac{\alpha^{(1)}}{n} \frac{\partial r(\mathbf{z}'_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{w}^0} \right\| + \|G(\mathbf{w}^0) - G(\mathbf{w}^0)\| \\ &\leq \frac{2\alpha^{(1)}}{n} \text{lip}(r \leftarrow \mathbf{w}); \end{aligned}$$

<sup>1</sup>In the cases of  $\mathbf{m}$  being a matrix, the matrix will be reshaped into a vector and the vector  $L_2$ -norm can then be used; this is equivalent to using the matrix Frobenius norm directly.

118  $t = 2$

$$\begin{aligned}
& \|\mathbf{m}^{(2)}(\mathbf{z}^n) - \mathbf{m}^{(2)}(\mathbf{z}^{n,i})\| \\
& \leq \left\| \frac{\alpha^{(2)}}{n} \frac{\partial r(\mathbf{z}_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{m}^{(1)}(\mathbf{z}^n)} - \frac{\alpha^{(2)}}{n} \frac{\partial r(\mathbf{z}'_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{m}^{(1)}(\mathbf{z}^{n,i})} \right\| + \\
& \quad \|G(\mathbf{m}^{(1)}(\mathbf{z}^n)) - G(\mathbf{m}^{(1)}(\mathbf{z}^{n,i}))\| \\
& \leq \frac{2\alpha^{(2)}}{n} \text{lip}(r \leftarrow \mathbf{w}) + \eta \frac{2\alpha^{(1)}}{n} \text{lip}(r \leftarrow \mathbf{w}) \\
& = \frac{2(\eta\alpha^{(1)} + \alpha^{(2)}) \text{lip}(r \leftarrow \mathbf{w})}{n};
\end{aligned}$$

119  $\vdots$   
120  $t = T$

$$\begin{aligned}
& \|\mathbf{m}^{(T)}(\mathbf{z}^n) - \mathbf{m}^{(T)}(\mathbf{z}^{n,i})\| \\
& \leq \left\| \frac{\alpha^{(T)}}{n} \frac{\partial r(\mathbf{z}_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{m}^{(T-1)}(\mathbf{z}^n)} - \frac{\alpha^{(T)}}{n} \frac{\partial r(\mathbf{z}'_i, \mathbf{w})}{\partial \mathbf{w}} \Big|_{\mathbf{m}^{(T-1)}(\mathbf{z}^{n,i})} \right\| + \\
& \quad \|G(\mathbf{m}^{(T-1)}(\mathbf{z}^n)) - G(\mathbf{m}^{(T-1)}(\mathbf{z}^{n,i}))\| \\
& \leq \frac{2\left(\sum_{t=1}^T \eta^{T-t} \alpha^{(t)}\right) \text{lip}(r \leftarrow \mathbf{w})}{n}.
\end{aligned}$$

121 (5) Derive the concentration inequality

$$\begin{aligned}
& |\mathbf{m}_{[q]}^{(T)}(\mathbf{z}^n) - \mathbf{m}_{[q]}^{(T)}(\mathbf{z}^{n,i})| \\
& \leq \|\mathbf{m}^{(T)}(\mathbf{z}^n) - \mathbf{m}^{(T)}(\mathbf{z}^{n,i})\| \\
& \leq \frac{2\left(\sum_{t=1}^T \eta^{T-t} \alpha^{(t)}\right) \text{lip}(r \leftarrow \mathbf{w})}{n} \\
& = \frac{C}{n},
\end{aligned}$$

122 where  $C = 2\left(\sum_{t=1}^T \eta^{T-t} \alpha^{(t)}\right) \text{lip}(r \leftarrow \mathbf{w})$ .

123 Based on Lemma B.1,  $\mathbf{m}_{[q]}^{(T)}(\mathbf{z}^n)$  can be bounded as

$$\begin{aligned}
\mathbb{P}_{\mathbf{z}^n} \left[ \left| \mathbf{m}_{[q]}^{(T)}(\mathbf{z}^n) - \mathbb{E}_{\mathbf{z}^n} \mathbf{m}_{[q]}^{(T)}(\mathbf{z}^n) \right| \leq \frac{\epsilon}{\sqrt{Q}} \right] & \geq 1 - 2 \exp\left(\frac{-2\epsilon^2}{Q \sum_{i=1}^n c_i^2}\right) \\
& = 1 - 2 \exp\left(\frac{-2\epsilon^2 n}{QC^2}\right).
\end{aligned}$$

124 Therefore,

$$\begin{aligned}
& \mathbb{P}_{\mathbf{z}^n} [\|\mathbf{m}^{(T)}(\mathbf{z}^n) - \mathbb{E}_{\mathbf{z}^n} \mathbf{m}^{(T)}(\mathbf{z}^n)\| \leq \epsilon] \\
(a) \quad & \geq \mathbb{P}_{\mathbf{z}^n} \left[ \bigcap_{q=1}^Q \left| \mathbf{m}_{[q]}^{(T)}(\mathbf{z}^n) - \mathbb{E}_{\mathbf{z}^n} \mathbf{m}_{[q]}^{(T)}(\mathbf{z}^n) \right| \leq \frac{\epsilon}{\sqrt{Q}} \right] \\
(b) \quad & \geq 1 - 2Q \exp\left(\frac{-2\epsilon^2 n}{QC^2}\right),
\end{aligned}$$

125 where inequality (a) is due the relationship between the events; inequality (b) is due to a Frechet  
126 inequality.  $\square$

### 127 C.5 Proof of Lemma 3

128 First, the definitions of Rademacher complexity, uniform convergence and covering number are  
129 introduced. Dudley's Integral Theorem that uses covering number to bound Rademacher complexity  
130 is also introduced. Then, by using the Lipschitz constant, the covering number of functional space is  
131 shown to be bounded by the covering number of parameter space. Finally, based on Dudley's Integral  
132 Theorem, Lemma 3 is shown.

133 **C.5.1 Preliminary**

134 **Definition C.1.** [5] Let  $\epsilon^n = \{\epsilon_1, \dots, \epsilon_n\}$  be i.i.d. random variables with  $P(\epsilon_i = 1) = P(\epsilon_i =$   
 135  $-1) = \frac{1}{2}$ .  $\mathbf{z}^n = \{z_1, \dots, z_n\}$  are i.i.d. samples. The *empirical Rademacher complexity* is defined as

$$\hat{\text{Rad}}_n(\mathcal{H}) = \mathbb{E}_{\epsilon^n} \left[ \max_{h \in \mathcal{H}} \frac{1}{n} \sum_i \epsilon_i h(z_i) \middle| \mathbf{z}^n \right];$$

136 and the *Rademacher complexity* is defined as

$$\text{Rad}(\mathcal{H}) = \mathbb{E}_{\mathbf{z}^n} \left[ \hat{\text{Rad}}_n(\mathcal{H}) \right].$$

137 **Theorem C.1.** [5] With probability at least  $1 - \delta$  the following bound holds:

$$R(h) - R_n(\mathbf{z}^n, h) \leq 2\hat{\text{Rad}}_n(\phi \circ \mathcal{H}) + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2n}},$$

138 where  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  denotes the loss function  $l(h(\mathbf{x}); y)$ ;  $\circ$  denotes the composition of functions.

139 **Lemma C.1.** [5] Let  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  be an  $L$ -Lipschitz function. Then, for any hypothesis set  $\mathcal{H}$  of  
 140 real-valued functions, *Talagrand's Lemma* indicates the following inequality holds:

$$\hat{\text{Rad}}_n(\phi \circ \mathcal{H}) \leq L\hat{\text{Rad}}_n(\mathcal{H}).$$

141 **Corollary C.1.** Suppose  $\text{lip}(r \leftarrow h) \leq L$ , then with probability at least  $1 - \delta$  the following bound  
 142 holds:

$$R(h) - R_n(\mathbf{z}^n, h) \leq 2L\hat{\text{Rad}}_n(\mathcal{H}) + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2n}}$$

143 *Proof.* Substituting the result of Lemma C.1 into Theorem C.1 gives the result.  $\square$

144 **Definition C.2.** [9] An  $\epsilon$ -cover of a subset  $\mathcal{U}$  of a metric space  $(\mathcal{V}, \rho)$  is a set  $\hat{\mathcal{U}} \subseteq \mathcal{U}$  such that for  
 145 each  $\mathbf{u} \in \mathcal{U}$  there is a  $\hat{\mathbf{u}} \in \hat{\mathcal{U}}$  such that  $\rho(\mathbf{u}, \hat{\mathbf{u}}) \leq \epsilon$ . The  $\epsilon$ -cover number of  $\mathcal{U}$  is

$$N(\epsilon, \mathcal{U}, \rho) = \min\{|\hat{\mathcal{U}}| : \hat{\mathcal{U}} \text{ is an } \epsilon\text{-cover of } \mathcal{U}\}.$$

146 The following theorem illustrates how to bound the covering number.

147 **Theorem C.2.** [9] Let  $\mathcal{U} \subseteq \mathcal{V} = \mathbb{R}^D$ . Then

$$\left(\frac{1}{\epsilon}\right)^D \frac{\text{vol}(\mathcal{U})}{\text{vol}(\mathcal{B})} \leq N(\epsilon, \mathcal{U}, \|\cdot\|) \leq \left(\frac{\text{vol}(\mathcal{U} + \frac{\epsilon}{2}\mathcal{B})}{\text{vol}(\frac{\epsilon}{2}\mathcal{B})}\right)$$

148 where  $+$  is the Minkovski sum,  $\mathcal{B}$  is the unit norm ball and  $\text{vol}$  indicates the volume of the set.

149 Remark: Consider  $\mathcal{U} \in \mathbb{R}^D$  with diameter  $\text{diam}(\mathcal{U})$ . Based on the last inequality, we have

$$N(\epsilon, \mathcal{U}, \|\cdot\|) \leq \left(\frac{\text{vol}(\mathcal{U} + \frac{\epsilon}{2}\mathcal{B})}{\text{vol}(\frac{\epsilon}{2}\mathcal{B})}\right) \leq \left(\frac{\text{diam}(\mathcal{U}) + \epsilon}{\epsilon}\right)^D = \left(1 + \frac{\text{diam}(\mathcal{U})}{\epsilon}\right)^D.$$

150 **Definition C.3.** Let  $\forall h_1, h_2 \in \mathcal{H}$  be two functions mapping  $z \in \mathcal{Z}$  into real value,  $\rho_{\mathcal{H}|\mathbf{z}^n}$  is defined  
 151 as follows:

$$\rho_{\mathcal{H}|\mathbf{z}^n}(h_1, h_2) = \sqrt{\frac{1}{n} \sum_{i=1}^n (h_1(z_i) - h_2(z_i))^2}.$$

152 **Theorem C.3.** [7] With metric  $\rho_{\mathcal{H}|\mathbf{z}^n}$  on  $\mathcal{H}$ , *Dudley's integral* indicates

$$\hat{\text{Rad}}_n(\mathcal{H}) \leq 12 \int_0^\infty \sqrt{\frac{\log N(\epsilon, \mathcal{H}, \rho_{\mathcal{H}|\mathbf{z}^n})}{n}} d\epsilon.$$

153 Dudley's integral bounds the empirical Rademacher complexity by the covering number of the  
 154 function space (with a metric based on the difference of the function value on  $n$  inputs).

155 **C.5.2 Bound of the covering number of functional space**

156 To start with, another definition of metric in function space is given as follows.

157 **Definition C.4.** A metric  $\rho_{\mathcal{H}_w}$  in parametric function space is defined as follows:

$$\rho_{\mathcal{H}_w}(h(\cdot; \mathbf{w}_1), h(\cdot; \mathbf{w}_2)) = \max_{\mathbf{x} \in \mathcal{X}} |h(\mathbf{x}; \mathbf{w}_1) - h(\mathbf{x}; \mathbf{w}_2)|. \quad (3)$$

158  $\text{lip}(h; \mathcal{H}_w \leftarrow \mathcal{W})$  will be written as  $\text{lip}(h \leftarrow \mathbf{w})$  if  $\mathcal{W}$  and  $\mathcal{H}_w$  are clear from the context:

$$\begin{aligned} \text{lip}(h \leftarrow \mathbf{w}) &= \max_{\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}, \mathbf{w}_1 \neq \mathbf{w}_2} \frac{\rho_{\mathcal{H}_w}(h(\cdot; \cdot, \mathbf{w}_1), h(\cdot; \cdot, \mathbf{w}_2))}{\rho_{\mathcal{W}}(\mathbf{w}_1, \mathbf{w}_2)} \\ &= \max_{\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}, \mathbf{w}_1 \neq \mathbf{w}_2, \mathbf{x}} \frac{|h(\mathbf{x}; \mathbf{w}_1) - h(\mathbf{x}; \mathbf{w}_2)|}{\|\mathbf{w}_1 - \mathbf{w}_2\|}. \end{aligned}$$

159 **Proposition C.2.** For all spaces of parametric functions  $\mathcal{H}_w$ ,  $\forall \epsilon, \forall \mathcal{H}$ ,

$$N(\epsilon, \mathcal{H}, \rho_{\mathcal{H}|z^n}) \leq N(\epsilon, \mathcal{H}, \rho_{\mathcal{H}_w}), \quad (4)$$

160 where  $\mathbf{w}$  denotes all parameters of the function,  $\rho_{\mathcal{H}|z^n}$  is defined in Definition C.3 and  $\rho_{\mathcal{H}_w}$  is defined  
161 in Definition C.4.

162 *Proof.* Let  $\{\hat{h}_1, \dots, \hat{h}_N\}$  be an  $\epsilon$ -covering set in  $\mathcal{H}_w$  with metric  $\rho_{\mathcal{H}_w}$ , then based on the definition  
163 of covering set,

$$\forall h \in \mathcal{H}, \min_j \rho_{\mathcal{H}_w}(h, \hat{h}_j) \leq \epsilon.$$

164 Based on the definitions of  $\rho_{\mathcal{H}|z^n}$  and  $\rho_{\mathcal{H}_w}$ , we have

$$\begin{aligned} \rho_{\mathcal{H}|z^n}(h, \hat{h}_j) &= \sqrt{\frac{1}{n} \sum_{i=1}^n (h(\mathbf{z}_i) - \hat{h}_j(\mathbf{z}_i))^2} \leq \sqrt{\frac{1}{n} \sum_{i=1}^n (\max_{\mathbf{z}} |h(\mathbf{z}) - \hat{h}_j(\mathbf{z})|)^2} \\ &= \sqrt{\frac{1}{n} \times n \times (\rho_{\mathcal{H}_w}(h, \hat{h}_j))^2} = \rho_{\mathcal{H}_w}(h, \hat{h}_j) \leq \epsilon. \end{aligned}$$

165 Therefore,  $\{\hat{h}_1, \dots, \hat{h}_N\}$  is also an  $\epsilon$ -covering set of  $\mathcal{H}_w$  with metric  $\rho_{\mathcal{H}|z^n}$  and

$$N(\epsilon, \mathcal{H}, \rho_{\mathcal{H}|z^n}) \leq |\{\hat{h}_1, \dots, \hat{h}_N\}| = N(\epsilon, \mathcal{H}, \rho_{\mathcal{H}_w}).$$

166  $\square$

167 **Corollary C.2.** The empirical Rademacher complexity can be bounded by the covering number with  
168 metric  $\rho_{\mathcal{H}_w}$  as follows:

$$\hat{\text{Rad}}_n(\mathcal{H}) \leq 12 \int_0^\infty \sqrt{\frac{\log N(\epsilon, \mathcal{H}, \rho_{\mathcal{H}_w})}{n}} d\epsilon.$$

169 *Proof.* Substituting the result of Proposition C.2 into Theorem C.3 gives the result.  $\square$

170 **Proposition C.3.** Let  $h(\mathbf{z}; \mathbf{w})$  be a parameterized function and  $\mathbf{w} \in \mathcal{W} \in \mathbb{R}^Q$ . Suppose  $\text{lip}(h \leftarrow$   
171  $\mathbf{w}) \leq L$ . Then,

$$N(\epsilon, \mathcal{H}_w, \rho_{\mathcal{H}_w}) \leq N(\epsilon/L, \mathcal{W}, \rho_{\mathcal{W}}) \leq \left(1 + \frac{\text{diam}(\mathcal{W})L}{\epsilon}\right)^Q.$$

172 *Proof.* The second inequality follows from Theorem C.2. We now show the first inequality. Let  
173  $\{\hat{\mathbf{w}}_1, \dots, \hat{\mathbf{w}}_N\}$  be an  $(\epsilon/L)$ -covering set in  $\mathcal{W}$ . Based on the definition of covering set,

$$\forall \mathbf{w} \in \mathcal{W}, \min_i \rho_{\mathcal{W}}(\mathbf{w}, \hat{\mathbf{w}}_i) \leq \epsilon/L.$$

174 Based on the definition of Lipschitz constant,

$$\forall h(\cdot; \mathbf{w}) \in \mathcal{H}_w, \min_i \rho_{\mathcal{H}_w}(h(\cdot; \mathbf{w}), h(\cdot; \hat{\mathbf{w}}_i)) \leq L \min_i \rho_{\mathcal{W}}(\mathbf{w}, \hat{\mathbf{w}}_i) \leq \epsilon.$$

175 Therefore,  $\{h(\cdot; \hat{\mathbf{w}}_1), \dots, h(\cdot; \hat{\mathbf{w}}_N)\}$  is a  $\epsilon$ -covering set of  $\mathcal{H}$  and

$$N(\epsilon, \mathcal{H}(\mathbf{w}), \rho_{\mathcal{H}_w}) \stackrel{(c)}{\leq} |\{h(\cdot; \hat{\mathbf{w}}_1), \dots, h(\cdot; \hat{\mathbf{w}}_N)\}| \stackrel{(d)}{\leq} |\{\hat{\mathbf{w}}_1, \dots, \hat{\mathbf{w}}_N\}| = N(\epsilon/L, \mathcal{W}, \rho_{\mathcal{W}}),$$

176 where inequality (c) is based on the definition of covering number; inequality (d) is due to the fact  
177 that  $h$  is a function.  $\square$

178 **C.5.3 Proof of Lemma 3**

179 *Proof.* Based on the result of Corollary C.2,

$$\begin{aligned}
& \widehat{\text{Rad}}_n(\mathcal{H}) \\
& \leq 12 \int_0^\infty \sqrt{\frac{\log N(\epsilon, \mathcal{H}, \rho_{\mathcal{H}_w})}{n}} d\epsilon \stackrel{(a)}{=} 12 \int_0^{LB} \sqrt{\frac{\log N(\epsilon, \mathcal{H}, \rho_{\mathcal{H}_w})}{n}} d\epsilon \\
& \stackrel{(b)}{\leq} \frac{12}{\sqrt{n}} \int_0^{LB} \sqrt{\log\left(1 + \frac{LB}{\epsilon}\right)^Q} d\epsilon \stackrel{(c)}{=} \frac{12LB}{\sqrt{n}} \int_0^1 \sqrt{Q \log\left(1 + \frac{1}{\epsilon'}\right)} d\epsilon' \\
& \stackrel{(d)}{\leq} 12LB \sqrt{\frac{Q}{n}} \int_0^1 \sqrt{\log(2/\epsilon')} d\epsilon' \stackrel{(e)}{=} 24LB \sqrt{\frac{Q}{n}} \int_0^{1/2} \sqrt{\log(1/\epsilon)} d\epsilon.
\end{aligned}$$

180 Equality (a) holds as the value of  $h$  is bounded by  $LB$ ; if  $\epsilon > LB$ , then  $\log N(\epsilon, \mathcal{H}, \rho_{\mathcal{H}_w}) = 0$ ;  
181 inequality (b) is based on Proposition C.3; equality (c) follows from variable substitution  $\epsilon' = \frac{\epsilon}{LB}$ ;  
182 inequality (d) is due to  $\epsilon' \in [0, 1]$ ; equality (e) follows from another variable substitution  $\epsilon = \frac{\epsilon'}{2}$ .

183 Then we calculate the integral

$$\begin{aligned}
& \int_0^{1/2} \sqrt{\log(1/\epsilon)} d\epsilon \\
& \stackrel{(a)}{=} \int_\infty^{\sqrt{\log 2}} y d(e^{-y^2}) \stackrel{(b)}{=} e^{-y^2} y|_\infty^{\sqrt{\log 2}} - \int_\infty^{\sqrt{\log 2}} e^{-y^2} dy \\
& = e^{-y^2} y|_\infty^{\sqrt{\log 2}} + \int_{\sqrt{\log 2}}^\infty e^{-y^2} dy \leq e^{-y^2} y|_\infty^{\sqrt{\log 2}} + \int_0^\infty e^{-y^2} dy \\
& = \frac{\sqrt{\log 2}}{2} + \frac{\sqrt{\pi}}{2},
\end{aligned}$$

184 where equality (a) is based on variable substitution  $y = \sqrt{\log(1/\epsilon)}$ , i.e.  $\epsilon = e^{-y^2}$  and equality (b) is  
185 based on integration by parts.

186 Therefore,

$$\begin{aligned}
\widehat{\text{Rad}}_n(\mathcal{H}) & \leq 24LB \sqrt{\frac{Q}{n}} \int_0^{1/2} \sqrt{\log(1/\epsilon)} d\epsilon \\
& \leq 24 \left( \frac{\sqrt{\log 2}}{2} + \frac{\sqrt{\pi}}{2} \right) LB \sqrt{\frac{Q}{n}} \\
& = CLB \sqrt{\frac{Q}{n}},
\end{aligned}$$

187 where  $C = 12(\sqrt{\log 2} + \sqrt{\pi})$ .

188 Finally, substituting the above bound of empirical Rademacher complexity into Corollary C.1 gives  
189 Lemma 3.  $\square$

190 **C.6 Proof of Theorem 3**

191 *Proof.* Let  $\text{ball}(E, \epsilon) := \text{ball}(\mathbb{E}_{z^n} \mathbf{m}^{(T)}(z^n), \epsilon)$  denote the ball with the center at  $\mathbb{E}_{z^n} \mathbf{m}^{(T)}(z^n)$   
192 and radius of  $\epsilon$ . Let  $L = \text{lip}(r \leftarrow \mathbf{w})$ . Based on Lemma 2, we have

$$\mathbb{P}_{z^n}[\mathbf{m}(z^n) \in \text{ball}(E, \epsilon)] \geq 1 - \delta_1, \tag{5}$$

193 where  $\delta_1 = 2Q \exp\left(\frac{-2\epsilon^2 n}{Q(2C_2)^2 L^2}\right)$ , that is  $\epsilon = C_2 L \sqrt{\frac{2Q}{n} \ln \frac{2Q}{\delta_1}}$ .

194 Based on the result of Lemma 3,

$$\mathbb{P}_{z^n} \left[ \max_{\mathbf{w} \in \text{ball}(E, \epsilon)} (R(\mathbf{w}) - R_n(z^n, \mathbf{w})) \leq CL(2\epsilon) \sqrt{\frac{Q}{n}} + \sqrt{\frac{\ln 1/\delta_2}{2n}} \right] \geq 1 - \delta_2.$$

195 Substituting  $\epsilon = C_2 L \sqrt{\frac{2Q}{n} \ln \frac{2Q}{\delta_1}} \leq C_2 L_1 L_l \sqrt{\frac{2Q}{n} \ln \frac{2Q}{\delta_1}}$  into the above formula, we have

$$\mathbb{P}_{\mathbf{z}^n} \left[ \max_{\mathbf{w} \in \text{ball}(E, \epsilon)} (R(\mathbf{w}) - R_n(\mathbf{z}^n, \mathbf{w})) \leq \epsilon' \right] \geq 1 - \delta_2, \quad (6)$$

196 where

$$\epsilon' = \frac{2CC_2L_1^2L_l^2Q\sqrt{2\ln(2Q/\delta_1)}}{n} + \sqrt{\frac{\ln(1/\delta_2)}{2n}}.$$

197 Based on Theorem 2, the final result is obtained by combining Eqs. 5,6 and setting  $C_1 = 2\sqrt{2}C$ :

$$\mathbb{P}_{\mathbf{z}^n} [R(\mathbf{m}(\mathbf{z}^n)) - R_n(\mathbf{z}^n, \mathbf{m}(\mathbf{z}^n)) \leq \epsilon] \geq 1 - \delta_1 - \delta_2.$$

198

□

## 199 D Lipschitz smoothness and updating equations of SMILE

200 For a classifier  $h$  with convex constraints on parameters, the parameter  $\mathbf{w}$  will be restricted to be  
 201 inside a convex set, as explained in Sec. D.1. Then based on Corollary A.1, a sufficient condition for  
 202 bounded  $\text{lip}(\frac{\partial h}{\partial \mathbf{w}} \leftarrow \mathbf{w})$  is to have finite values of the first and second partial derivatives.

### 203 D.1 Equivalence between constrained optimization and the use of regularization terms

204 Let us review two optimization problems.

205 Problem 1:

$$\min_{\mathbf{w}} R_n(\mathbf{z}_n, h_{\mathbf{w}}) \quad \text{s.t. } \mathcal{P}(\mathbf{w}) \leq C;$$

206 Problem 2:

$$\min_{\mathbf{w}} R_n(\mathbf{z}_n, h_{\mathbf{w}}) + \lambda \mathcal{P}(\mathbf{w}).$$

207 The Lagrange function of Problem 1 is

$$\mathcal{L}(\mathbf{w}, u) = R_n(\mathbf{z}_n, h_{\mathbf{w}}) + u(\mathcal{P}(\mathbf{w}) - C), \quad u \geq 0,$$

208 where  $u$  is the Lagrangian multiplier.

209 For Problem 1, the (KKT) necessary conditions imply

$$\text{Condition 1} \quad \frac{\partial R_n(\mathbf{z}_n, h_{\mathbf{w}})}{\partial \mathbf{w}} + u \frac{\partial \mathcal{P}(\mathbf{w})}{\partial \mathbf{w}} = 0;$$

$$\text{Condition 2} \quad u(\mathcal{P}(\mathbf{w}) - C) = 0.$$

210 For Problem 2, the necessary condition implies

$$\frac{\partial R_n(\mathbf{z}_n, h_{\mathbf{w}})}{\partial \mathbf{w}} + \lambda \frac{\partial \mathcal{P}(\mathbf{w})}{\partial \mathbf{w}} = 0.$$

211 Suppose  $\mathbf{w}_1^*$  and  $\mu^*$  satisfy the necessary condition of Problem 1. Setting  $\lambda = \mu^*$ , we can see that  
 212  $\mathbf{w}_1^*$  satisfies for the necessary condition of Problem 2. Suppose  $\mathbf{w}_2^*$  satisfies the necessary condition  
 213 of Problem 2. Setting  $\mu = \lambda$  and  $C = \mathcal{P}(\mathbf{w}_2^*)$ , we can see that Condition 1 and Condition 2 of  
 214 Problem 1 are satisfied, so  $\mathbf{w}_2^*$  satisfies the necessary condition of Problem 1 as well. Based on the  
 215 above results, the necessary conditions of Problem 1 and Problem 2 are equivalent.

216 Meanwhile, when the regularization term in Problem 2 is a convex function, the equivalent Problem 1  
 217 constrains  $\mathbf{w}$  inside the set of  $\{\mathbf{w} | \mathcal{P}(\mathbf{w}) \leq C\}$ , which is a convex set [1].

### 218 D.2 First partial derivative of SMILE classifier

219 The first partial derivatives of the classifier (Eq. 9) are as follows:

$$\frac{\partial h(\mathbf{x}; \mathcal{W})}{\partial \mathbf{r}_j^+} = -\exp(-\|\mathbf{L}\mathbf{x} - \mathbf{r}_j^+\|^2)(2\mathbf{r}_j^+ - 2\mathbf{L}\mathbf{x})$$

$$\frac{\partial h(\mathbf{x}; \mathcal{W})}{\partial \mathbf{r}_j^-} = \exp(-\|\mathbf{L}\mathbf{x} - \mathbf{r}_j^-\|^2)(2\mathbf{r}_j^- - 2\mathbf{L}\mathbf{x})$$

$$\begin{aligned} \frac{\partial h(\mathbf{x}; \mathcal{W})}{\partial \mathbf{L}_{[a,b]}} &= -\sum_j 2(\mathbf{L}\mathbf{x} - \mathbf{r}_j^+)_{[a]}\mathbf{x}_{[b]} \exp(-\|\mathbf{L}\mathbf{x} - \mathbf{r}_j^+\|^2) \\ &\quad + \sum_k 2(\mathbf{L}\mathbf{x} - \mathbf{r}_k^-)_{[a]}\mathbf{x}_{[b]} \exp(-\|\mathbf{L}\mathbf{x} - \mathbf{r}_k^-\|^2), \end{aligned}$$

220 where  $L_{[a,b]}$  denotes the  $a$ th row and  $b$ th column element of matrix  $L$  and  $x_{[a]}$  denotes the  $a$ th  
 221 element of the vector  $\mathbf{x}$ ;  $(L\mathbf{x} - \mathbf{r})_{[a]} = \sum_i L_{[a,i]}x_{[i]} - r_{[a]}$ .

222  $\frac{\partial h(\mathbf{x}; \mathcal{W})}{\partial r_j^+}$  and  $\frac{\partial h(\mathbf{x}; \mathcal{W})}{\partial r_j^-}$  are bounded by  $2 \text{diam}(\mathbf{r}_a^-) + 2 \text{diam}(L) \text{diam}(\mathbf{x})$ ;  $\frac{\partial h(\mathbf{x}; \mathcal{W})}{\partial L}$  is bounded  
 223 by  $4m(\text{diam}(\mathbf{r}) + \text{diam}(L) \text{diam}(\mathbf{x})) \text{diam}(\mathbf{x})$ , where  $m$  denotes the number of representative  
 224 instances. All first partial derivatives have finite values as long as  $\text{diam}(L)$ ,  $\text{diam}(\mathbf{x})$  and  $\text{diam}(\mathbf{r})$   
 225 are bounded.

### 226 D.3 Second partial derivative of SMILE classifier

227 The second partial derivatives are as follows:

$$\begin{aligned} \frac{\partial^2 h(\mathbf{x}; \mathcal{W})}{\partial r_i^{+2}} &= 4 \exp(-\|L\mathbf{x} - \mathbf{r}_j^+\|^2) (\mathbf{r}_j^+ - L\mathbf{x})(\mathbf{r}_j^+ - L\mathbf{x})^T - 2 \exp(-\|L\mathbf{x} - \mathbf{r}_j^+\|^2) \mathbf{I}; \\ \frac{\partial^2 h(\mathbf{x}; \mathcal{W})}{\partial r_j^{-2}} &= -4 \exp(-\|L\mathbf{x} - \mathbf{r}_k^-\|^2) (\mathbf{r}_j^- - L\mathbf{x})(\mathbf{r}_j^- - L\mathbf{x})^T + 2 \exp(-\|L\mathbf{x} - \mathbf{r}_k^-\|^2) \mathbf{I}; \\ \frac{\partial^2 h(\mathbf{x}; \mathcal{W})}{\partial L_{[a,b]}^2} &= \sum_j 4(L\mathbf{x} - \mathbf{r}_j^+)_{[a]}^2 x_{[b]}^2 \exp(-\|L\mathbf{x} - \mathbf{r}_j^+\|^2) - 2 \sum_j x_{[b]}^2 \exp(-\|L\mathbf{x} - \mathbf{r}_j^+\|^2) \\ &\quad - \sum_k 4(L\mathbf{x} - \mathbf{r}_k^-)_{[a]}^2 x_{[b]}^2 \exp(-\|L\mathbf{x} - \mathbf{r}_k^-\|^2) + 2 \sum_k x_{[b]}^2 \exp(-\|L\mathbf{x} - \mathbf{r}_k^-\|^2), \end{aligned}$$

228 where  $\mathbf{I}$  is the identity matrix. All second partial derivatives have finite values as long as  $\text{diam}(L)$ ,  
 229  $\text{diam}(\mathbf{x})$  and  $\text{diam}(\mathbf{r})$  are bounded.

### 230 D.4 Updating equations of SMILE

231 The updating equations of SMILE are as follows:

$$\begin{aligned} \mathbf{r}_j^{+,t+1} &= \mathbf{r}_j^{+,t} - 2\lambda\alpha \mathbf{r}_j^{+,t} + \frac{\alpha}{n} \sum_{i=1}^n y_i l'(y_i h(\mathbf{x}_i; \mathcal{W})) \exp(-\|L\mathbf{x}_i - \mathbf{r}_j^+\|^2) (2\mathbf{r}_j^+ - 2L\mathbf{x}_i) |_{\mathcal{W}^t}; \\ \mathbf{r}_k^{-,t+1} &= \mathbf{r}_k^{-,t} - 2\lambda\alpha \mathbf{r}_k^{-,t} - \frac{\alpha}{n} \sum_{i=1}^n y_i l'(y_i h(\mathbf{x}_i; \mathcal{W})) \exp(-\|L\mathbf{x}_i - \mathbf{r}_k^-\|^2) (2\mathbf{r}_k^- - 2L\mathbf{x}_i) |_{\mathcal{W}^t}; \\ \mathbf{L}^{t+1} &= \mathbf{L}^t - 2\lambda\alpha \mathbf{L}^t + \frac{\alpha}{n} \sum_{i=1}^n y_i l'(y_i h(\mathbf{x}_i; \mathcal{W})) \sum_j \exp(-\|L\mathbf{x}_i - \mathbf{r}_j^+\|^2) 2(L\mathbf{x}_i - \mathbf{r}_j^+) \mathbf{x}_i^T |_{\mathcal{W}^t} \\ &\quad - \frac{\alpha}{n} \sum_{i=1}^n y_i l'(y_i h(\mathbf{x}_i; \mathcal{W})) \sum_k \exp(-\|L\mathbf{x}_i - \mathbf{r}_k^-\|^2) 2(L\mathbf{x}_i - \mathbf{r}_k^-) \mathbf{x}_i^T |_{\mathcal{W}^t}. \end{aligned}$$

## 232 E Data description

233 Table 1 lists information on sample  
 234 size and feature dimension, as well as  
 235 the source of studied datasets.

Table 1: Data description

Dataset	Source	# instances	# features
Australian	UCI	690	14
Cancer	UCI	699	9
Climate	UCI	540	18
Credit	UCI	653	15
German	UCI	1000	24
Haberman	UCI	306	3
Heart	UCI	270	13
ILPD	UCI	583	10
Liver	UCI	345	6
Pima	UCI	768	8
Ringnorm	Delve	7400	20
Twonorm	Delve	7400	20

236 **References**

- 237 [1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press,  
238 2004.
- 239 [2] David Haussler. Decision theoretic generalizations of the PAC model for neural net and other  
240 learning applications. *Information and Computation*, pages 78–150, 1992.
- 241 [3] John H Hubbard and Barbara Burke Hubbard. *Vector calculus, linear algebra, and differential*  
242 *forms: a unified approach*. Matrix Editions, 2015.
- 243 [4] H Quang Minh and Thomas Hofmann. Learning over compact metric spaces. In *International*  
244 *Conference on Computational Learning Theory*, pages 239–254. Springer, 2004.
- 245 [5] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*.  
246 MIT press, 2012.
- 247 [6] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to*  
248 *Algorithms*. Cambridge University Press, 2014.
- 249 [7] N Srebro and K Sridharan. Note on refined Dudley integral covering number bound, 2010.
- 250 [8] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142,  
251 1984.
- 252 [9] M. J. Wainwright. *High-dimensional Statistics: A Non-asymptotic Viewpoint*. Cambridge  
253 University Press, 2019.
- 254 [10] Nik Weaver. *Lipschitz Algebras*. World Scientific, 1999.