

Theory of Learning and Kernel methods (CS5470)

Summary Notes

August 10, 2020

Contents

Contents	1
1 Introduction	3
2 Examples of Statistical Consistency	7
2.1 Finite-sized Model+Bounded Loss	8
2.2 Models with restricted Capacity	10
2.2.1 Linear Classifier Model with 0-1 Loss	13
2.2.2 Bounded Linear Model with Hinge-loss and bounded data . .	14
2.2.3 Kernelized Linear Models	18
3 Model Selection and Bayes Consistency	25
3.1 Introduction	25
3.2 Model Selection	26
3.2.1 Finitely-Sized Model Class	26
3.2.2 Infinitely-sized Model Class: example of Bayes Consistency . .	27

Chapter 1

Introduction

Towards the goal of formally studying the problem of “(machine) learning”, we began by defining¹ few terms that facilitate mimicking the process of learning in humans:

Datapoint/Input Intuitively, this is a placeholder that represents a single “input” or “stimulus” given to a human (learner). We denote it by x and the set of all possible inputs by $\mathcal{X}$, called the **input space**. Formally, $\mathcal{X}$ can be any arbitrary set. To begin with, we consider $\mathcal{X} = \mathbb{R}^n$ (Euclidean vectors).

Label/Target/Output This is a placeholder for the expected “output” or “response” or “action” from the human (learner) given a stimulus. We denote it by y and the set of all possible labels by $\mathcal{Y}$, called the **label space (or output space)**. To begin with, we restrict $\mathcal{Y}$ to be a discrete set of two (uncomparable) objects. In such a case, the learning problem is known as “**binary classification**”.

(Unknown) Concept Intuitively, this represents the abstract concept that captures the relation between input and output. Formally, concept can be any valid joint distribution² over³ $\mathcal{X} \times \mathcal{Y}$. We denote this unknown distribution by F_{XY} .

We then wanted to focus on a particular learning setting that is easy to analyze. Our choice of the learning setting goes by the name “**Supervised Induc-**

¹Definitions of terms/notation are hence-forth highlighted in *blue*.

²With the usual Borel measure over $\mathbb{R}^n$, distribution function is the familiar “cumulative distribution function (cdf)”. Recall that, cdf exists and characterizes a random variable, irrespective of whether it is continuous or discrete or neither.

³one may have to make $\mathcal{X}, \mathcal{Y}$ as measurable spaces for this. Atleast, in our case, when $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y}$ is a finite set, this is well-understood.

tive Learning (SIL)”. In SIL, the learner is presented a set of **training examples** $\mathcal{D}_m = \{(x_1, y_1) \dots (x_m, y_m)\}$, known as the **training set**. The learner attempts to “learn” a function $\hat{g}_m : \mathcal{X} \mapsto \mathcal{Y}$ that incurs the least⁴ expected⁵ “error”⁶. The algorithm used for arriving at $\hat{g}_m$ is called as the **training/learning** algorithm.

We then went on to define error formally. In ML, error is measured using what is called a **loss function**, which is any function $l : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$. We will begin with the so called **0-1 loss** that is given by $l(y, g(x)) = 1$ if $y \neq g(x)$, and 0 otherwise. With this, it is immediate to define **Bayes optimal function**⁷:

$$(1.1) \quad g^{**} \in \arg \min_{g \in \mathcal{U}} R[g],$$

where $R[g] \equiv \mathbb{E}[l(Y, g(X))]$ (expectation taken under F_{XY}), and $\mathcal{U} \equiv \{h \mid h : \mathcal{X} \mapsto \mathcal{Y}\}$ ⁸. R is known as **(true) risk functional** and $R[g]$ hence denotes the (true) risk of employing the prediction function g in the application at hand. Though g^{**} is the skyline function, it can never be computed, in general. Hence the goal in SIL can only be to find a $\hat{g}_m$ that can (in some sense) approximate g^{**} .

We then defined a **learning algorithm (SIL algorithm)**, $\mathcal{A}$, as any function that takes as input (any) training set (of any size m) and outputs a function ($\hat{g}_m : \mathcal{X} \mapsto \mathcal{Y}$). The function output by an algorithm is known as the **prediction function** (associated with the particular training set given). We say an algorithm $\mathcal{A}$ is **Bayes consistent**, under the given loss (l), iff the (infinite) sequence of prediction functions $\hat{g}_1, \dots, \hat{g}_n, \dots$ associated with random⁹ training sets of sizes $1, \dots, n, \dots$, satisfies the following property:

$$(1.2) \quad \{R[\hat{g}_n]\} \xrightarrow{p} R[g^{**}]$$

for any¹¹ F_{XY} . Here $\xrightarrow{p}$ denotes asymptotic convergence in probability.

Again, we further restricted the SIL setting to a (non-probabilistic) model-based one. In such a setting, the prediction function must belong to a pre-defined

⁴or atleast “close” to the least with high probability

⁵Expectation taken under the unknown distribution, F_{XY} .

⁶We will define error (loss) shortly.

⁷By definition, this need not be unique i.e., the argmin might be a non-singleton set. The handwritten notes circulated includes a simplified expression for Bayes optimal in a special case.

⁸More formally, one has to restrict $\mathcal{U}$ to only measurable functions from $\mathcal{X}$ to $\mathcal{Y}$. Also, let's assume $\mathcal{U}, F_{XY}$ are such that the expectation is always well-defined.

⁹Apologies for the slack in notation for $\hat{g}_m$. It is random variable if associated with a random training set¹⁰ and it is a fixed function if associated with a given training set.

¹¹Some textbooks define this as universal Bayes consistency instead of Bayes consistency. And, if the above happens for only a particular distribution, F , then they say it is Bayes consistent wrt. F .

(given) set of functions (from $\mathcal{X}$ to $\mathcal{Y}$), called the [model](#)¹². We denote a model by $\mathcal{G}$ and elements in $\mathcal{G}$ by g . Needless to say, now the skyline prediction function also needs to be re-defined:

$$(1.3) \quad g^* \in \arg \min_{g \in \mathcal{G}} R[g].$$

We say a (model-based) SIL algorithm is statistically consistent, under the given loss (l), iff the (infinite) sequence of prediction functions $\hat{g}_1 \in \mathcal{G}, \dots, \hat{g}_n \in \mathcal{G}, \dots$ associated with random training sets of sizes $1, \dots, n, \dots$, satisfies the following property:

$$(1.4) \quad \{R[\hat{g}_n]\} \xrightarrow{p} R[g^*] \quad (\text{as } n \rightarrow \infty),$$

for any F_{XY} .

While the above consistency notions are asymptotic, since in real-world only finite number of samples are available, one defines the following notions of error incurred by a prediction function g :

1. [Estimation error](#) with g (under loss l and model $\mathcal{G}$) is defined as $R[g] - R[g^*]$.
2. [Approximation error](#) (or model error) with g is defined as $R[g^*] - R[g^{**}]$.
3. [Generalization error](#) with g is defined as $R[g] - R[g^{**}]$ i.e., generalization error is the sum of estimation error and approximation error.

We now begin with example(s) of statistically consistent algorithms and later on see how to extend them to achieve Bayes consistency.

¹²In probabilistic modeling, model is a set of distributions. Each distribution is such that a prediction function can be induced from it.

Chapter 2

Examples of Statistical Consistency

We motivated a natural algorithm for SIL, henceforth known as [Empirical Risk Minimization \(ERM\)](#), which simply minimizes the average loss over the training set. The expression for average loss over m training examples, $\mathcal{D}_m = \{(x_1, y_1), \dots, (x_m, y_m)\}$, is given by, $\hat{R}_m[g] \equiv \frac{\sum_{i=1}^m l(y_i, g(x_i))}{m}$, where $\hat{R}_m$ is popular by the name [Empirical Risk functional](#). We denote the function in the model that minimizes this empirical risk by¹ $\hat{g}_m$.

It was immediate that ERM is statistically consistent with any loss function if the model has only one function (trivial, useless case). A useful, but easy to analyze case is that of Models with finite elements, detailed in section 2.1. Before that, the following provides a useful sufficiency criteria for ERM (statistical) consistency:

Note that, 0 is

$$\begin{aligned} &\leq R[\hat{g}_m] - R[g^*], && (\because \text{defn. of } g^*), \\ &= \left(R[\hat{g}_m] - \hat{R}_m[\hat{g}_m] \right) + \left(\hat{R}_m[\hat{g}_m] - \hat{R}_m[g^*] \right) + \left(\hat{R}_m[g^*] - R[g^*] \right), && (\because \text{add-subtract terms}), \\ &\leq \left(R[\hat{g}_m] - \hat{R}_m[\hat{g}_m] \right) + \left(\hat{R}_m[g^*] - R[g^*] \right), && (\because \text{defn. of } \hat{g}_m), \\ &\leq \max_{g \in \mathcal{G}} \left(R[g] - \hat{R}_m[g] \right) + \left(\hat{R}_m[g^*] - R[g^*] \right) . \end{aligned}$$

Since we know that the second expression in the last row of the above goes to

¹Again, we abuse notation for $\hat{R}_m, \hat{g}_m$ i.e., $\hat{R}_m, \hat{g}_m$ are fixed functions if training set is given, and $\hat{R}_m, \hat{g}_m$ are random variables if a random training set is considered.

zero in probability, the sufficiency condition for ERM (statistical) consistency is:

$$(2.1) \quad \left\{ \max_{g \in \mathcal{G}} \left(R[g] - \hat{R}_m[g] \right) \right\} \xrightarrow{p} 0 \quad (\text{as } m \rightarrow \infty).$$

The above sufficiency criteria is popularly known as the “one-sided uniform convergence” criteria.

2.1 Finite-sized Model+Bounded Loss

The key technical result of this section appears as theorem 2.2 in Mohri et al. [2012]. Also, interested students are encouraged to read the entire section 2 in this book. The version of this result derived in the lectures appears below.

We shall now show that (2.1) holds for models that are finitely-sized when the loss function is bounded. Using Boole’s inequality we have:

$$(2.2) \quad P \left[\max_{g \in \mathcal{G}} \left(R[g] - \hat{R}_m[g] \right) > \epsilon \right] \leq \sum_{g \in \mathcal{G}} P \left[R[g] - \hat{R}_m[g] > \epsilon \right].$$

Now we require to bound probabilities involving deviations of average of iid random variables from its mean. [Chernoff bounding technique](#) [Chernoff, 1952], is a general technique which provides a bound for probability of a linear function of independent random variables deviating from its true mean. The key steps in this technique are²:

- $P \left[R[g] - \hat{R}_m[g] > \epsilon \right] = P \left[e^{s(R[g] - \hat{R}_m[g])} > e^{s\epsilon} \right] \forall s > 0,$
- Applying Markov inequality gives $\text{RHS} \leq e^{-s\epsilon} \mathbb{E}[e^{s(R[g] - \hat{R}_m[g])}] \forall s > 0,$
- Use the fact that the random variables³ $L_1(g), L_2(g), \dots, L_m(g)$ are independent (infact iid): $\text{RHS} \leq e^{-s\epsilon} \prod_{i=1}^m \mathbb{E}[e^{\frac{s}{m}(\mathbb{E}[L_i(g)] - L_i(g))}]$
- Use the [Hoeffding lemma](#), detailed in http://en.wikipedia.org/wiki/Hoeffding%27s_lemma, to bound the moment generating function (mgf) of the mean zero and finitely supported random variable $\mathbb{E}[L_i(g)] - L_i(g)$ (finite support is true

²Note that the technique is generic and when applied with different partial information about the involving random variables and the function combining them, one gets different bounds. We will shortly see another bound called McDiarmid’s inequality which follows most of these basic steps. You can also refer sec.5.2 in Schölkopf and Smola [2002] for detailed derivation (for case $|calF| = 1$). Here we provide the version with the relevant random variables for the present context.

³We denote the random variable (X_i, Y_i) by $\mathbf{Z}_i$ and the random variable $l(Y_i, g(X_i))$ by $L_i(g)$.

whenever the loss function is bounded, which in particular is true with zero-one loss): $\text{RHS} \leq e^{-s\epsilon} e^{\frac{s^2}{2m}}$ (for 0-1 loss⁴).

- Finally, choose the best s (by minimizing the bound on RHS): $\text{RHS} \leq e^{-\frac{1}{2}m\epsilon^2}$

In summary, we have: $P \left[R[g] - \hat{R}_m[g] > \epsilon \right] \leq e^{-\frac{1}{2}m\epsilon^2}$. Plugging this in (2.2), we have that $P \left[\max_{g \in \mathcal{G}} \left(R[g] - \hat{R}_m[g] \right) > \epsilon \right] \leq |\mathcal{G}| e^{-\frac{1}{2}m\epsilon^2}$. This bounding first of all shows that the probability term in question which is sandwiched between zero and $|\mathcal{G}| e^{-\frac{1}{2}m\epsilon^2}$ goes to zero as $m \rightarrow \infty$; confirming that ERM is consistent in finite $|\mathcal{G}|$ case⁵. Re-writing the bound by denoting $\delta = |\mathcal{G}| e^{-\frac{1}{2}m\epsilon^2}$ gives:

with probability atleast $1 - \delta$,

$$(2.3) \quad R[g] \leq \hat{R}_m[g] + \sqrt{\frac{2}{m} \log \left(\frac{|\mathcal{G}|}{\delta} \right)} \quad \forall g \in \mathcal{G}.$$

Inequalities of such type are called as VC-type inequalities⁶. Sometimes they are also referred to as learning bounds. Note that the learning bound holds uniformly for all the candidate functions in $\mathcal{G}$, including the empirical risk minimizers. Infact, ever since the one-sided uniform convergence criteria was written, all results indeed hold for every candidate function.

Interestingly the learning bound gives an upper-bound on the risk (the quantity we want to minimize) that involves terms that can be computed based on $\mathcal{D}$ and $\mathcal{G}$. Hence such bounds provide computable (upper) bounds on the performance (risk) of g obtained with an induction principle like ERM⁷.

Finally, note that the above derivation also shows that with probability atleast $1 - \delta$, we have, $R[g^*] \leq \hat{R}_m[g^*] + \sqrt{\frac{2}{m} \log \left(\frac{1}{\delta} \right)}$. Combining this learning bound with that in 2.3 using union-bound (Boole's inequality) gives: with probability atleast $1 - \delta$,

$$(2.4) \quad R[\hat{g}_m] \leq R[g^*] + \sqrt{\frac{2}{m} \log \left(\frac{2|\mathcal{G}|}{\delta} \right)} + \sqrt{\frac{2}{m} \log \left(\frac{2}{\delta} \right)}.$$

⁴For 0-1 loss, $0 \leq L_i(g) \leq 1$, hence $-1 \leq \mathbb{E}[L_i(g)] - L_i(g) \leq 1$.

⁵Note that the analysis is very similar in the countable case. It is the uncountable case which calls for a different analysis. Nevertheless at a later stage we will clarify why countable case is similar to the finite case.

⁶As they were popularized by Vapnik and Chervonenkis.

⁷We commented on the play between $|\mathcal{G}|, m, \delta$ and the tightness of the bound.

In other words, with high probability (atleast $1 - \delta$), the estimation error with ERM (and 0-1 loss with finitely-sized models) cannot exceed $\sqrt{\frac{2}{m} \log \left(\frac{2|\mathcal{G}|}{\delta} \right)} + \sqrt{\frac{2}{m} \log \left(\frac{2}{\delta} \right)}$.

We now move on to the more interesting, nevertheless complicated, case of models containing infinite candidate functions.

2.2 Models with restricted Capacity

In this case one cannot resort to the Boole's inequality (2.2) and one needs⁸ to directly focus on the random variable $h(Z_1, \dots, Z_m) = \max_{g \in \mathcal{G}} R[g] - \hat{R}_m[g]$. We noted that h is a function of iid random variables and moreover satisfies the bounded difference property⁹. Hence one can employ the [McDiarmid's inequality \[McDiarmid, 1989\]](#) to bound probability of high deviations of h from its mean. [Refer \[www.cs.berkeley.edu/~bartlett/courses/281b-sp06/bdddifff.pdf\]\(http://www.cs.berkeley.edu/~bartlett/courses/281b-sp06/bdddifff.pdf\) for an easy proof of the McDiarmid inequality and the definition of bounded difference property.](#) With this we have that with probability atleast $1 - \delta$,

$$(2.5) \quad R[g] \leq \hat{R}_m[g] + \mathbb{E} \left[\max_{g \in \mathcal{G}} R[g] - \hat{R}_m[g] \right] + \sqrt{\frac{1}{2m} \log \left(\frac{1}{\delta} \right)}, \quad \forall g \in \mathcal{G}$$

The equation holds for losses which vary between 0 and 1 (like 0-1 loss). Needless to say, a similar statement can be written for any bounded loss function.

We noted that the expectation in the RHS above represents how big a function class is and hence the VC-type inequality in the general $\mathcal{G}$ case is very similar to that in the finite case (2.3). In order that the bound is useful we wanted to further bound the expectation term (which is unknown). Following are the steps towards this goal:

Ghost Samples: $\mathbb{E} \left[\max_{g \in \mathcal{G}} R[g] - \hat{R}_m[g] \right] = \mathbb{E} \left[\max_{g \in \mathcal{G}} \mathbb{E} \left[\hat{R}'_m[g] \right] - \hat{R}_m[g] \right]$. Here $\hat{R}'_m[g] = \frac{1}{m} \sum_{i=1}^m l(Y'_i, g(X'_i))$ represents the empirical risk with g evaluated

⁸In lectures, we mentioned how intuitively the notion of Growth function (defined later) can be employed in this case. Interested students please refer to theorem 4.1 in <http://www.shivani-agarwal.net/Teaching/E0370/Aug-2011/Lectures/3.pdf> for the formal details. Nevertheless, we choose an alternate methodology, involving Rademacher complexity rather than growth function, here that gives more generic and slightly tighter bounds. Infact, later on we link it up with Growth functions too.

⁹We commented that bounded difference is indeed the key property satisfied by sample mean too and is responsible for its concentration around the true mean.

on a set of m iid samples $Z'_1, \dots, Z'_m$ (called ghost samples; $Z'_i \equiv (X'_i, Y'_i)$) which are independent of the training set.

Max. and Expectation interchange: Since maximum of sum/integral is less than or equal to sum/integral of maxima, we have¹⁰: $\mathbb{E} \left[\max_{g \in \mathcal{G}} \mathbb{E} \left[\hat{R}'_m[g] \right] - \hat{R}_m[g] \right] \leq \mathbb{E} \left[\max_{g \in \mathcal{G}} \hat{R}'_m[g] - \hat{R}_m[g] \right]$. Note that the final expectation is wrt. both Z_i and Z'_i for all i . We defined $D_m(\mathcal{H}) \equiv \mathbb{E} \left[\max_{h \in \mathcal{H}} \frac{2}{m} \sum_{i=1}^{\frac{m}{2}} h(Z_i) - \frac{2}{m} \sum_{i=\frac{m}{2}+1}^m h(Z_i) \right]$ as the maximum discrepancy of $\mathcal{H}$ (at m). If we define $\mathcal{G} \circ l \equiv \{l(\cdot, g(\cdot)) \mid g \in \mathcal{G}\}$, as the loss-class induced by the model $\mathcal{G}$ and loss l , then it is easy to see that the quantity of interest (appearing as RHS of the inequality immediately above) is $D_{2m}(\mathcal{G} \circ l)$. Though this is a valid measure of “complexity of a model”, we wished to further upper bound this with a more comprehensive measure that looks at all possible splits (instead of a half-split alone in D_m). Towards this goal, we introduce new random variables $\sigma_1, \dots, \sigma_m$, called Rademacher variables, which are iid with distribution: $P[\sigma_i = 1] = 0.5, P[\sigma_i = -1] = 0.5$:

Rademacher variables: We have, $\mathbb{E} \left[\max_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^m (l(Y'_i, g(X'_i)) - l(Y_i, g(X_i))) \right]$

$$= \mathbb{E} \left[\max_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^m \sigma_i (l(Y'_i, g(X'_i)) - l(Y_i, g(X_i))) \right].$$

This equality is true because the distribution of $(l(Y'_i, g(X'_i)) - l(Y_i, g(X_i)))$ is symmetrical (wrt. 0). Note that the expectation in the last expression is wrt. all random variables i.e., $Z_i, Z'_i, \sigma_i, \forall i$.

Again, max. and sum inequality: $\mathbb{E} \left[\max_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^m \sigma_i (l(Y'_i, g(X'_i)) - l(Y_i, g(X_i))) \right] \leq \mathbb{E} \left[\max_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^m \sigma_i (l(Y'_i, g(X'_i))) \right] + \mathbb{E} \left[\max_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^m -\sigma_i (l(Y_i, g(X_i))) \right]$. Again, because of symmetry, this is further equal to $2\mathbb{E} \left[\max_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^m \sigma_i l(Y_i, g(X_i)) \right]$. This motivates the following definition: Rademacher complexity of a function class $\mathcal{H}$ (at m) is defined as:

$$\mathcal{R}_m(\mathcal{H}) \equiv \mathbb{E} \left[\max_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i h(Z_i) \right],$$

where the expectation is over the random variables $Z_i, \sigma_i, \forall i$. With this notation, the expectation in the final upper bound above is nothing but $\mathcal{R}_m(\mathcal{G} \circ l)$. Note that in addition to the interpretation that the Rademacher complexity measures the expected “maximum discrepancy” across all random splits, it can

¹⁰This explanation is perhaps more apt than the contrived Jensen’s inequality argument presented in books.

also be interpreted as a measure of the expected maximum correlation with (Rademacher) noise¹¹. Since Rademacher complexity involves the unknown distribution, it can never be computed. So we wish to further bound it in terms of the corresponding empirical average: the Rademacher complexity conditioned on the training examples is called the **conditional Rademacher complexity**: $\hat{\mathcal{R}}_m(\mathcal{H}) \equiv \mathbb{E} \left[\max_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^m \sigma_i g(Z_i) \mid Z_1, \dots, Z_m \right]$. Note that unlike $\mathcal{R}_m$, the quantity $\hat{\mathcal{R}}_m$ can be computed (given the training set). Hence we would like to have a bound in terms of $\hat{\mathcal{R}}_m$ rather than $\mathcal{R}_m$:

McDiarmid Inequality: It is easy to see that the function $h(Z_1, \dots, Z_m) = \hat{\mathcal{R}}(\mathcal{G} \circ l)$ satisfies bounded difference property and hence application of McDiarmid's inequality¹² gives with probability at least $1 - \delta$:

$$(2.6) \quad \mathcal{R}_m(\mathcal{G} \circ l) = \mathbb{E} \left[\hat{\mathcal{R}}_m(\mathcal{G} \circ l) \right] \leq \hat{\mathcal{R}}_m(\mathcal{G} \circ l) + \sqrt{\frac{1}{2m} \log \left(\frac{1}{\delta} \right)}.$$

Union bound: Combining equations (2.5) and (2.6) with a union bound (Boole's inequality), and using all the above inequalities, we have with probability at least $1 - \delta$:

$$(2.7) \quad R[g] \leq \hat{R}_m[g] + 2\hat{\mathcal{R}}_m(\mathcal{G} \circ l) + 3\sqrt{\frac{1}{2m} \log \left(\frac{2}{\delta} \right)}, \quad \forall g \in \mathcal{G}$$

Hence a sufficiency condition for ERM being (statistically) consistent is of course $\hat{\mathcal{R}}_m(\mathcal{G} \circ l) \xrightarrow{p} 0$ as $m \rightarrow \infty$. Another not very different sufficiency condition is $\mathcal{R}_m(\mathcal{G} \circ l) \rightarrow 0$ as $m \rightarrow \infty$. Clearly this does not happen with $\mathcal{G}$ being the set of all (measurable) functions as in that case $\hat{\mathcal{R}} = 0.5$ (assuming 0-1 loss). This establishes the statement that (non-trivial) statistical consistency may not be possible unless the function class is restricted in its complexity (as measured by Rademacher complexity). In the subsequent section we look at some nice loss-classes where the Rademacher complexity decays leading to ERM consistency.

¹¹Motivated by this interpretation, one can also talk about correlation with Gaussian noise (in place of Rademacher noise). This leads to the notion of Gaussian complexity. Interested students may look at Bartlett and Mendelson [2002] for details on the different complexity measures and their relations.

¹²Again, the inequality is written with 0-1 loss mind. Similar expression for any bounded loss can be written.

2.2.1 Linear Classifier Model with 0-1 Loss

We began with the case of linear classifier model, i.e., $\mathcal{G} = \{g \mid \exists w \in \mathbb{R}^n \ni g(x) = \text{sign}(w^\top x) \forall x \in \mathcal{X} \subset \mathbb{R}^n\}$, and the $l = 0-1$ loss. We then proved Lemma 3.1 in Mohri et al. [2012] that gives $\mathcal{R}_m(\mathcal{G} \circ l) = 0.5\mathcal{R}_m(\mathcal{G})$. Infact this lemma holds for any model who elements are functions with binary output $+1/-1$, but loss is 0-1.

We then proved Massart’s lemma (see theorem 3.3 and its corollaries in Mohri et al. [2012]), which helped us bound the Rademacher average in this case with an expression involving the **growth function**: $\mathcal{R}_m(\mathcal{G}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{G}}(m)}{m}}$, where $\Pi_{\mathcal{G}} : \mathbb{N} \mapsto \mathbb{N}$ is the Growth function of $\mathcal{G}$ and is given by $\Pi_{\mathcal{G}}(m) \equiv \max_{x_1, \dots, x_m} |\{(g(x_1), \dots, g(x_m)) \mid g \in \mathcal{G}\}|$. It simply is the maximum number of distinct function values the functions in the model induce on m datapoints.

For the case of linear classifiers, we gave a simple upper bound¹³ of $2^{n+1} \binom{m}{n}$ on the Growth function, where n is the dimensionality of the data. Using this and the Sterling’s approximation, we have that $\mathcal{R}_m(\mathcal{G}) \leq O\left(\sqrt{\frac{n \log m}{m}}\right)$.

The good news is that the final bound shows that ERM is statistically consistent in this case. And the rate of convergence, in terms of m , is also not too worse compared to the finite-sized model case. However, the bad news is that the complexity is now a function of (grows with) the dimensionality of the data. This is not desirable because, typically, one expects to “learn” well given richer descriptions of the data (more dimensionality), but inherently this model penalizes such data! This is called the “curse of dimensionality” in machine learning literature. The subsequent section presents a slight modification to this loss-class that breaks this curse.

Also, 0-1 loss is not attractive from a practical perspective: the ERM problem (with the linear classifier model and 0-1 loss) is computationally hard¹⁴. This is another reason for employing alternative losses, for e.g., hinge-loss in the subsequent section.

¹³Generalization of this bound for any function class that has finite growth function is called Sauer’s lemma. Refer theorem 3.5 in Mohri et al. [2012] and <http://www.cs.berkeley.edu/~bartlett/courses/281b-sp08/16.pdf> for proof of the same. This leads to the learning bound (3.31) in Mohri et al. [2012] (Please read pages 31-48 in Mohri et al. [2012] for derivation of this bound.). This bound involves the so called **VC-dimension** of the function class (Refer defn. 3.4 for definition and examples 3.1-3.5 for example computations of VC-dimension).

¹⁴Infact a more comprehensive statement can be made: refer ? for details.

2.2.2 Bounded Linear Model with Hinge-loss and bounded data

Given bounded data, say $\max_{x \in \mathcal{X}} \|x\| = R$, we intuitively argued out that the growth function depends on the margin of classification rather than the dimensionality; thus potentially breaking the curse of dimensionality. We then picked the loss-class accordingly: we choose the model as the (norm-bounded) linear model:

$$\mathcal{G}_W = \{g \mid \exists w \in \mathbb{R}^n \ni g(x) = w^\top x \ \forall x \in \mathcal{X} \subset \mathbb{R}^n, \|w\| \leq W\}$$

and the loss function as the hinge-loss¹⁵: $l(y, g(x)) \equiv \max(0, 1 - yg(x))$. With such a loss-class it is clear that the minimum “gap” or “margin” between any two zero-loss incurring datapoints of opposite labels is greater than $\frac{2}{W}$ i.e., we are encouraging “fat-classification” or classification with large-margin.

We begin by noting that bounds analogous to (2.5)-(2.7) can be written for any bounded loss. In the present case, we have that $l(y, g(x)) \leq 1 + |w^\top x| \leq 1 + WR$, by Cauchy-Schwartz inequality. Hence we have with probability atleast $1 - \delta$:

$$(2.8) \quad R[g] \leq \hat{R}_m[g] + 2\hat{\mathcal{R}}(\mathcal{G}_W \circ l) + 3(1 + WR) \sqrt{\frac{1}{2m} \log \left(\frac{2}{\delta} \right)}, \ \forall g \in \mathcal{G}$$

We came up with an upper bound on the conditional Rademacher average in this case¹⁶ (we assume things as and when necessary):

Contraction Lemma: Let us denote the hinge-loss $l(y, g(x))$ by $\psi(yg(x)) = \max(0, 1 - yg(x))$. Then, we have:

$$\hat{\mathcal{R}}(\mathcal{G}_W \circ l) \equiv \mathbb{E} \left[\max_{\|w\| \leq W} \frac{1}{m} \sum_{i=1}^m \sigma_i \psi(y_i w^\top x_i) \right] \leq \mathbb{E} \left[\max_{\|w\| \leq W} \frac{1}{m} \sum_{i=1}^m \sigma_i y_i w^\top x_i \right].$$

This inequality follows from the contraction lemma [Ledoux and Talagrand, 1991] (refer Lemma5 in Meir and Zhang [2003] for a simple proof), because ψ is a Lipschitz continuous function¹⁷ with Lipschitz constant as unity.

Cauchy-Schwartz Inequality:

$$\mathbb{E} \left[\max_{\|w\| \leq W} \frac{1}{m} \sum_{i=1}^m \sigma_i y_i w^\top x_i \right] \leq \frac{W}{m} \mathbb{E} \left[\left\| \sum_{i=1}^m \sigma_i y_i x_i \right\| \right] = \frac{W}{m} \mathbb{E} \left[\sqrt{\hat{\sigma}^\top K \hat{\sigma}} \right],$$

¹⁵Note that we are still talking about binary classification and hence $\mathcal{Y} = \{-1, +1\}$

¹⁶The derivation presented here is based on the proof of theorem 24 in Lanckriet et al. [2004]

¹⁷A function f is said to be Lipschitz continuous with Lipschitz constant L iff $|f(x) - f(y)| \leq L\|x - y\| \ \forall x, y \in \text{dom}(f)$.

where $\hat{\sigma}$ is the vector with entries as $\sigma_i y_i$ and K is the matrix of all possible dot products: $(i, j)^{th}$ entry in K is $K_{ij} = x_i^\top x_j$. Such a matrix is called a [gram matrix](#). So K is the gram matrix of the training datapoints.

Jensen's Inequality: $\frac{W}{m} \mathbb{E} \left[\sqrt{\hat{\sigma}^\top K \hat{\sigma}} \right] \leq \frac{W}{m} \sqrt{\mathbb{E} [\hat{\sigma}^\top K \hat{\sigma}]}$ and this is equal to $\frac{W}{m} \sqrt{\text{trace}(K)}$, as σ_i are iid with mean zero and variance unity¹⁸.

Radius bound: Recall that we assumed that the input space $\mathcal{X}$ is bounded i.e., there exists an R such that $\|x\| \leq R \forall x \in \mathcal{X}$. With this assumption one obtains the following radius-margin bound:

$$(2.9) \quad \hat{\mathcal{R}}(\mathcal{G}_W \circ l) \leq \frac{WR}{\sqrt{m}},$$

which indeed goes to zero as $m \rightarrow \infty$.

Hence ERM should be consistent in this case. Infact, using (2.9), (2.8), we have with probability atleast $1 - \delta$:

$$(2.10) \quad R[g] \leq \hat{R}_m[g] + 2 \frac{WR}{\sqrt{m}} + 3(1 + WR) \sqrt{\frac{1}{2m} \log \left(\frac{2}{\delta} \right)}, \forall g \in \mathcal{G}$$

Using similar learning theory bounds Vapnik [Vapnik, 1998] proposed a optimization formalism that implements the ERM principle. This is the well celebrated formulation of [SVMs \(Support Vector Machines\)](#), which is the subject of discussion in section 2.2.2.

On passing we made an important comment that the function class we started with had no “curse of dimensionality”, as the expression for guaranteed risk is (not very loosely) upper-bounded by an expression independent of the input space dimensionality. We also commented that, in early 1990s (time of birth of SVMs), such non-cursed loss-classes were not known¹⁹. Another elegant feature of this loss-class and the Radius-Margin bound is that now we have an explicit single parameter that can control the complexity of the loss-class, namely W .

We end this section by pointing out that in the context of this loss-class itself, if one additionally assumes that the loss incurred is never non-zero, then it is clear that one is talking about “fat” linear classifiers (or gap tolerant classifiers) that insist that data points lie on either side of the fat slab. In this case, under the assumption that

¹⁸Trace of matrix M is sum of its diagonal entries

¹⁹Recall that the VC-dim of set of all linear classifiers is $d + 1$, where d is dimensionality of the input space; and is indeed NOT independent of dimensionality.

the data is bounded, one can upper bound the VC dimension again by something which is nearly dimensionally independent and dependent on the margin. Please refer to theorem 8.4 and its proof in Vapnik [1998]. Also, refer to sections 3.3, 7.1 in Burges [1998] for more detailed proof.

Support Vector Machines (SVMs)

Motivated by the result that ERM is consistent, one can look for a linear function which solves the following problem:

$$(2.11) \quad \begin{aligned} \min_{w \in \mathbb{R}^n} \quad & \sum_{i=1}^m l(x_i, y_i, w), \\ \text{s.t.} \quad & \|w\| \leq W \end{aligned}$$

One may use the truncated hinge loss or any upper bound of it. For eg. hinge loss. The advantage with hinge-loss is it is convex²⁰, whereas the truncated hinge-loss is not. With hinge loss (2.11) can be written as:

$$(2.12) \quad \begin{aligned} \min_{w \in \mathbb{R}^n} \quad & \sum_{i=1}^m \max(0, 1 - y_i w^\top x_i), \\ \text{s.t.} \quad & \|w\| \leq W \end{aligned}$$

The above problem is convex (and hence can be solved efficiently). Infact it can be posed as a Second-Order Cone Program (SOCP)²¹, once the objective is turned linear: we used a standard trick of introducing additional variables ξ_i such that $\xi_i \geq \max(0, 1 - y_i w^\top x_i)$. This gives:

$$(2.13) \quad \begin{aligned} \min_{w \in \mathbb{R}^n} \quad & \sum_{i=1}^m \xi_i, \\ \text{s.t.} \quad & \|w\| \leq W, \quad \xi_i \geq 0, \quad y_i w^\top x_i \geq 1 - \xi_i. \end{aligned}$$

Infact problems of the form (2.11) have been studied in optimization theory. Most common example is with the case of square-loss (regression problem). The term in the objective measures the fit of the model to the data, while the constraint “regularizes” the model. Such a regularization is known as Ivanov regularization. Moreover, regularization problems can be written in two more equivalent forms:

Tikhonov regularization:

$$(2.14) \quad \min_{w \in \mathbb{R}^n} \|w\| + C \sum_{i=1}^m l(x_i, y_i, w),$$

²⁰One may also re-derive the bounds for hinge-loss case, which would lead to similar expressions and results.

²¹refer <http://stanford.edu/~boyd/papers/socp.html>.

where C is a parameter (plays a role similar to W). Here the interpretation is fit the model to the data while regularizing it. C controls the trade-off between data fit and regularization. Some also refer to such a form as “Regularized risk minimization” (which we have shown is equivalent to ERM). Here regularized risk refers to the weighted sum of the regularizer and empirical risk.

Morozov regularization:

$$(2.15) \quad \begin{aligned} \min_{w \in \mathbb{R}^n} \quad & \|w\|, \\ \text{s.t.} \quad & \sum_{i=1}^m l(x_i, y_i, w) \leq A, \end{aligned}$$

where A is a parameter similar to C and W . Here the interpretation maximally regularize the model while data fit is under certain tolerance. A is a bound on the (empirical) error of data fit.

The Tikhonov regularized version with hinge-loss was used by Cortes and Vapnik [1995] and published as SVMs (only difference being $0.5\|w\|^2$ is used instead of $\|w\|$ as the regularizer):

$$(2.16) \quad \begin{aligned} \min_{w \in \mathbb{R}^n} \quad & \frac{1}{2}\|w\|^2 + C \sum_{i=1}^m \xi_i, \\ \text{s.t.} \quad & \xi_i \geq 0, \quad y_i w^\top x_i \geq 1 - \xi_i. \end{aligned}$$

The squared version of the regularizer was used to obtain a nice convex Quadratic Program (as above), for which highly efficient off-the-shelf solvers exist.

The Morozov regularized version (with squared-regularizer, hinge-loss and $A = 0$ i.e., no empirical error) was used in a preliminary paper before SVM [Boser et al., 1992] and leads to what usually is known as the hard-margin SVM:

$$(2.17) \quad \begin{aligned} \min_{w \in \mathbb{R}^n} \quad & \frac{1}{2}\|w\|^2, \\ \text{s.t.} \quad & y_i w^\top x_i \geq 1. \end{aligned}$$

Please read Burges [1998], which is an excellent tutorial on SVMs. Here we tried to cover things not covered there (including learning theory results). We next provide an insight into the specialty of the solution with the SVM problem that will be helpful in our analysis later on.

Note that the geometric interpretation of (2.17) is that of maximally separating two set of points. It is well known that this problem is equivalent to minimizing distance between convex hulls of the two sets of points²². Infact, the normal to the maximally separating hyperplane (i.e., w) will be in the direction of line joining the

²²Infact, this equivalence drives all duality principles in optimization. Refer notes at <http://www.cse.iitb.ac.in/saketh/teaching/cs709.html> for details.

two minimum distant points in the convex hulls. From this it is immediate that $w = \sum_{i=1}^m \alpha_i x_i$. Infact, later on we will (rigorously) prove a more generic statement under the name “Representer theorem” — which says (loosely) any “SVM-kind” of problem (i.e., norm-regularized linear fit problem) has a solution of the form $w = \sum_{i=1}^m \alpha_i x_i$ i.e., the solution is a linear combination of the training datapoints. Moreover, the name “Support Vector” is also motivated from this duality result: from the above argument it is also clear that many α s can be zero at optimality and hence the solution is a linear combination of few important examples called “support vectors”. Will fill-in more details as and when required.

We ended this discussion by writing down the optimization problems corresponding to various loss functions and functions classes: $\mathcal{F}_W^l$ with square-loss is known as ridge-regression Hoerl and Kennard [1970] or regularized least-squares or min. norm least squares²³. $\mathcal{F}_W^l$ with square-hinge loss is referred to as l_2 -SVM. $\mathcal{F}_W^l$ with ϵ -insensitive loss is called as Support Vector Regression Smola and Schölkopf [2004]. $\mathcal{F}_W^1$ with square-loss is called as LASSO Tibshirani [1996]. $\mathcal{F}_W^1$ with hinge-loss is called as l_1 -regularized SVM etc.

2.2.3 Kernelized Linear Models

We began by noting that the entire derivation in section 2.2.2 can be generalized²⁴ with the following model: $\mathcal{G}_W^\phi = \{g \mid \exists w \in \mathcal{H} \ni g(x) = \langle w, \phi(x) \rangle_{\mathcal{H}} \forall x \in \mathcal{X}, \|w\|_{\mathcal{H}} \leq W\}$, where $\mathcal{H}$ is a given Hilbert space²⁵, and²⁶ $\phi : \mathcal{X} \mapsto \mathcal{H}$ is a feature mapping from input space to the Hilbert space.

The key advantage is that now the model is not restricted to have only linear functions (over x). Indeed this is useful as Bayes optimal (for the given problem) need not be well-approximated by linear functions. In other words, if one desires an algorithm to be Bayes consistent, then the model-class, $\mathcal{G}_W$ for various W , cannot be restricted to those containing linear functions alone.

However, one disadvantage with this approach is that the dimensionality of ϕ grows exponentially with the degree of polynomial in x . For e.g., if $x \in \mathbb{R}^n$ and $\langle w, \phi(x) \rangle$ must be polynomial upto degree d , then $\phi(x)$ must contain all monomials upto degree d , which are $\binom{n+d-1}{d}$ in number! Indeed this number is exponential in d . While it is not a big issue theoretically as there is no “curse of dimensionality”, this

²³Refer to the limit defn. of Pseudo-inverse.

²⁴Through examples of affine and quadratic functions, we noted that non-linear functions in input space $\mathcal{X}$ are nothing but linear functions in a suitable (non-linearly) transformed space $\phi(\mathcal{X})$. e.g. $f(x) = ax_1^2 + bx_2^2 + \sqrt{2}cx_1x_2 = [a \ b \ c]^\top \phi(x)$, $\phi(x) = [x_1^2 \ x_2^2 \ \sqrt{2}x_1x_2]^\top$ (here $x = [x_1 \ x_2]^\top \in \mathbb{R}^2$). We also noted this is the case with all polynomial functions.

²⁵Please refer https://en.wikipedia.org/wiki/Hilbert_space for definition.

²⁶ $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, $\| \cdot \|_{\mathcal{H}}$ denote the inner-product and norm in the Hilbert space $\mathcal{H}$.

poses serious computational issues. More gravely, if one indeed desires a model-class suitable for Bayes consistency, then perhaps no restriction on d must exist. In other words, the most interesting case is perhaps when the $\mathcal{H}$ is infinite-dimensional. How does one even represent w efficiently (as it is infinite-dimensional) and then solve the ERM problem?

The critical issue of representation is solved thanks to the so-called Representer theorem:

Theorem 2.2.1. *Consider the problem of ERM with the model $\mathcal{G}_W^\phi$:*

$$(2.18) \quad \begin{aligned} \min_{w \in \mathcal{H}} \quad & \frac{1}{m} \sum_{i=1}^m l(y_i, \langle w, \phi(x_i) \rangle_{\mathcal{H}}), \\ \text{s.t.} \quad & \|w\|_{\mathcal{H}} \leq W. \end{aligned}$$

Then an optimal solution of the ERM problem of the form: $w = \sum_{i=1}^m \alpha_i \phi(x_i)$ exists for some $\alpha_i \in \mathbb{R}$. Needless to say, the same statement holds for the Tikhonov and Morozov forms of the above Ivanov ERM problem.

Refer theorem 4.2 in Schölkopf and Smola [2002] for details of the proof.

Using representer theorem, it is easy to see that (2.18) can be re-written as:

$$(2.19) \quad \begin{aligned} \min_{\alpha \in \mathbb{R}^m} \quad & \frac{1}{m} \sum_{i=1}^m l(y_i, \sum_{j=1}^m \alpha_j \langle \phi(x_j), \phi(x_i) \rangle_{\mathcal{H}}), \\ \text{s.t.} \quad & \alpha^\top G_m^\phi \alpha \leq W^2, \end{aligned}$$

where G_m^ϕ is the [gram matrix](#) over the points $\phi(x_1), \dots, \phi(x_m)$, i.e., $(i, j)^{th}$ entry of G_m^ϕ is $\langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$. Since gram-matrices are always positive-semi-definite, (2.19) is a convex problem whenever l is convex (in its second argument). In summary, the ERM problem (and the evaluation of prediction function) can be solved efficiently (even in infinite-dimensional Hilbert spaces) if inner-products $\langle \phi(x_j), \phi(x_i) \rangle_{\mathcal{H}}$ can be computed efficiently between any $x_i \in \mathcal{X}, x_j \in \mathcal{X}$.

Mathematicians studied inner-products of the form $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$ and showed that one does not need to explicitly construct $\phi, \mathcal{H}$ whenever k satisfies certain properties. This goes by the subject name “kernels”.

We began exploring kernels by asking what properties does a function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ need to satisfy in order to represent an inner-product of the form appearing above. It was immediate that the matrices of kernel evaluations over a given set of points $x_1, \dots, x_m$ will be the gram-matrix over the mapped points $\phi(x_1), \dots, \phi(x_m)$. Hence a necessary condition is all such matrices are positive-semi-definite.

With this motivation we define the following: [Given an input space \$\mathcal{X}\$ \(need not be Euclidean; infact need not be a vector space\), a kernel is any function \$k :\$](#)

$\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ satisfying i) symmetry: $x, z \in \mathcal{X} \Rightarrow k(x, z) = k(z, x)$ and ii) Non-negativity: $x_1, \dots, x_m \in \mathcal{X} \Rightarrow \alpha^\top G_m^k \alpha \geq 0 \forall \alpha$, where G_m^k is the matrix with ij^{th} entry as $k(x_i, x_j)$ i.e., it is the matrix of all possible kernel evaluations on the given set of m points. In other words, $G_m^k \succeq 0$ ²⁷.

One can now prove the following crucial theorem [Schölkopf and Smola, 2002]:

Theorem 2.2.2. *Consider an input space $\mathcal{X}$ and a valid kernel k over it. Then there exists a Hilbert space $\mathcal{H}_k$ and a feature map $\phi_k : \mathcal{X} \rightarrow \mathcal{H}_k$ such that the kernel evaluation of any two datapoints in the input space, i.e., $k(x, z)$, is equal to the inner product of those two datapoints in the feature space, i.e., $\langle \phi_k(x), \phi_k(z) \rangle_{\mathcal{H}_k}$. In other words, $k(x, z) = \langle \phi_k(x), \phi_k(z) \rangle_{\mathcal{H}_k}$.*

Refer section 2.2.2 in Schölkopf and Smola [2002] for a proof of the same²⁸.

Note that this proof not only shows existence of a Hilbert space, but a special one where $f \in \mathcal{H}_k \Rightarrow f(x) = \langle f(\cdot), k(x, \cdot) \rangle_{\mathcal{H}_k}$. This special Hilbert space is called the **Reproducing Kernel Hilbert Space (RKHS)** associated with k . Obviously there may be several space and mappings satisfying this criteria. Refer to theorem 2.10 and proposition 2.12 in Schölkopf and Smola [2002] for an alternate Hilbert space, actually an l_2 space, construction..

Now all this development is useful, only if we show some examples of positive kernels. Before giving examples let's look at some operations that preserve kernality, which come in handy: i) inner-products are kernels ii) conic combination of kernels is a kernel iii) product of kernels is a kernel iv) limit of a sequence of kernels (if it exists) is a kernel v) Normalized version of a kernel²⁹ is a kernel. Refer section 13.1 in Schölkopf and Smola [2002] for details. Though these results are simple to prove we argued that from application perspective they are far reaching: consider an application involving multi-modal data (say, video, audio, text modes) and suppose kernels for video, audio and text data are given. By linearly combining products of such kernels, one can obtain (non-trivial) feature representations for the multi-modal data!

²⁷The symbol $M \succeq 0$ means that the matrix M is positive semi-definite (psd) $M \succeq 0 \Leftrightarrow x^\top M x \geq 0 \forall x$ and M is symmetric. Some textbooks may prefer to define psd matrices (not necessarily symmetric) satisfying this non-negativity condition

²⁸Justification of (2.31) in Schölkopf and Smola [2002] needs to be done as we did in lecture rather than as done in Schölkopf and Smola [2002]. Basically we need Cauchy-Schwartz inequality to hold for any two functions in Hilbert space rather than for kernels alone. In lecture we showed that this is indeed the case. Also in the lecture we gave a nice justification for the choice of the feature map, which is at the heart of the proof. We said that representing an object by its similarities with all other objects is the most obvious representation (and infact the richest representation).

²⁹If k is a kernel, then $\bar{k}(x_1, x_2) \equiv \frac{k(x_1, x_2)}{\sqrt{k(x_1, x_1)}\sqrt{k(x_2, x_2)}}$ is its normalized version.

Now, we introduce the [kernelized linear model](#):

$$\mathcal{G}_W^k \equiv \{g \mid \exists w \in \mathcal{H}_k \ni g(x) = \langle w, \phi_k(x) \rangle_{\mathcal{H}_k} \forall x \in \mathcal{X}, \|w\|_{\mathcal{H}_k} \leq W\},$$

where $\mathcal{H}_k$ is the RKHS associated with k and $\phi_k : \mathcal{X} \mapsto \mathcal{H}_k$ is the canonical feature mapping from input space to the RKHS defined by $\phi_k(x)(\cdot) \equiv k(x, \cdot)$. From representer theorem it follows that the relevant functions in this space essentially as of the form $\sum_i \alpha_i k(x_i, x)$.

We showed many examples of kernels over Euclidean input spaces as well as non-euclidean ones. Here is the standard recipe for generating the examples:

1. Start with the inner-product kernel: $k_1(x_1, x_2) \equiv \langle x_1, x_2 \rangle$. Here, we assume, $\mathcal{X}$ is a subset in an inner-product space. For e.g.,
 - (a) $\mathcal{X} = \mathbb{R}^n$, then the general form for inner-product $\langle x_1, x_2 \rangle$ is $x_1^\top M x_2$, $M \succ 0^{30}$. M will be the kernel-hyper-parameter.
 - (b) $\mathcal{X}$ is set of all random variables (over $\mathbb{R}$). Then $\langle X, Y \rangle \equiv \mathbb{E}[XY]$ defines a valid inner-product between random variables X, Y .
 - (c) $\mathcal{X}$ is set of all probability density functions (over $\mathbb{R}$). Then $\langle f_X, f_Y \rangle \equiv \mathbb{E}_X[f_Y(X)] = \mathbb{E}_Y[f_X(Y)]$ (expected likelihood) defines a valid inner-product between densities f_X, f_Y .

The inner-product kernels are also known as [linear-kernels](#) because the functions induced³¹ by it are linear functions.

2. We form polynomials using k_1 : $k_2(x_1, x_2) \equiv (c + k_1(x_1, x_2))^d$. Because summation, product preserve kernality, k_2 is a valid kernel. These are called [polynomial kernels](#) as the functions induced are polynomials. Here $c > 0, d \in \mathbb{N}$ (and perhaps³² $M \succeq 0$) are hyper-parameters.
3. $k_3(x_1, x_2) \equiv e^{k_1(x_1, x_2)}$. Because summation, product, limits preserve kernality, k_3 is a valid kernel. The normalized version of k_3 is k_4 . This is called [Gaussian kernel](#) as the functions induced are linear combinations of Gaussian functions.

The other trick to generate kernels is to simply evaluate any of the above kernels over some function of the input rather than the input. For e.g., $k_4(x_1, x_2) \equiv k(h(x_1), h(x_2))$, where h is any arbitrary function whose range is the space over which

³⁰ $M \succ 0$ means M is (symmetric) positive-definite matrix.

³¹ As noted earlier, functions induced by k are of the form $\sum_i \alpha_i k(x_i, x)$.

³² Note that $M \succ 0$ is necessary for $k(x, y) = x^\top M y$ to be an inner-product. But $M \succeq 0$ is sufficient (and $M \succ 0$ is NOT necessary) for the same being a kernel.

k is defined. Examples of kernels constructed using this trick are the probability-product kernels presented in Jebara et al. [2004]. We then gave an example of another interesting kernel over distributions: Song et al. [2013].

We then discussed some desirable properties of kernels from a ML point of view:

1. **Normalized kernels:** This is because the radius of any data³³ with normalized kernel will be unity.
2. **Positive-definite kernel:** A kernel is positive-definite, iff all gram-matrices induced by it are positive-definite. We gave a simple linear-algebra proof for positive-definiteness of (un-normalized as well as normalized) Gaussian kernel.
3. **Universal kernel:** A kernel k over an input space $\mathcal{X}$ is said to be a universal kernel (for that space) iff the function class induced by the kernel is dense in the set of all continuous functions $\mathcal{C}(\mathcal{X})$. Then by Luzin’s theorem Folland [1996] it follows that functions induced by k will approximate any Bayes optimal. Using Stone-Weierstrass theorem [Rudin, 1976] we verified that the (un-normalized as well as normalized) Gaussian kernel is a universal kernel. Refer theorem 1 in Steinwart [2001] for details. On passing, we note the following paper Christmann and Steinwart [2010], which provides examples of universal kernels over non-Euclidean spaces.

We conclude our discussion on kernels by noting another recipe for their construction. Here $\mathcal{X}$ is arbitrary (need not be a subset of a Hilbert space) but we assume domain knowledge:

1. Ask the domain expert to provide a function that measures (pair-wise) similarity³⁴ between inputs in $\mathcal{X}$.
2. if the similarity function is a kernel, use it as k_1 . Else search for a kernel that is closest to the similarity function and use it as k_1 .
3. Now repeat steps 2-3 in the previous recipe with this k_1 .

Note that this is a more generic way of constructing kernels and the goal is not limited to inducing non-linear functions³⁵.

³³In context of radius-margin bound

³⁴Many diverse domain experts feel that providing similarity function is easier/intuitive/well-suited than providing a Euclidean vector representation for inputs.

³⁵In fact, there is no notion of “linear functions” in this situation.

Finally, we can write down the ERM problem using kernelized model as follows:

$$(2.20) \quad \begin{aligned} \min_{\alpha \in \mathbb{R}^m} \quad & \sum_{i=1}^m l \left(y_i \sum_{j=1}^m \alpha_j k(x_i, x_j) \right), \\ \text{s.t.} \quad & \sqrt{\alpha^\top G_k \alpha} \leq W. \end{aligned}$$

Here G_k is the matrix of all kernel evaluations on the training points. Again, this has same convexity and other properties as (2.19). In the subsequent chapter, we focus on Bayes consistency.

Chapter 3

Model Selection and Bayes Consistency

3.1 Introduction

So far, we have discussed methods to select the “best” g from a given model $\mathcal{G}$.

Recall that Empirical Risk Minimization (ERM) assures statistical consistency, i.e., as $m \rightarrow \infty$, $\{R[\hat{g}_m]\} \rightarrow R[g^*]$. In other words, the estimation error i.e., $R[\hat{g}_m] - R[g^*]$ asymptotically decays to zero. However, for proving statistical consistency, we placed restrictions on the model $\mathcal{G}$: Rademacher complexity of corresponding loss-class must decay to zero. On the other hand, the following observations are true:

- If $\mathcal{G}$ is large, then the Approximation Error i.e., $R[g^*] - R[g^{**}]$ is less and,
- If $\mathcal{G}$ is small (and the Rademacher complexity is small), then the Estimation Error i.e., $R[\hat{g}_m] - R[g^*]$ is less

This is a trade-off in the sense that, if we want to approach closer to the Bayes’ optimal classifier, we need a larger model, but doing so makes the ERM convergence harder. While this trade-off is important to perform model selection (topic in subsequent section), it is clear that unless one considers infinite number of (restricted) models, one cannot guarantee Bayes consistency (topic in section 3.2.2). We first deal with the simpler problem of model selection with finite number of models and then graduate to the case of infinite models leading to Bayes consistency.

3.2 Model Selection

We begin with some definitions: A model class is a collection of models. Given a model class, the goal in model selection is to pick the “best” model among the ones in the model class. Needless to say, parameter selection (via, say, ERM) is then performed in that “best” model. The obvious measure of badness of a model is the generalization error (Approximation-error+Estimation-error) that shall be incurred by employing the model (at the very least). In the subsequent section, we present a series of algorithms that perform model selection in a finitely-sized class.

3.2.1 Finitely-Sized Model Class

Let the number of models be d . We begin with the following algorithms that merely are statistically consistent:

ERM Perform ERM (using the given dataset) in each model to obtain d candidates. Then pick the one among those d with the least ER (computed using the same dataset). The model from which this final candidate is picked is declared as the “best” model. This final candidate is used as the prediction function.

Validation Partition the given dataset into two: call one as the training set and the other as the validation set. Perform ERM using the training set (alone) in each model to obtain d candidates. Then pick the one among those d with the least [Validation-set error](#), i.e., the empirical risk computed using the validation set. The model from which this final candidate is picked is declared as the “best” model. This final candidate is used as the prediction function.

k-fold Cross-Validation Partition the given dataset into k parts. Run the previous algorithm (Validation) for k times, wherein the i^{th} iteration, the i^{th} part is used as the Validation set and the rest $k - 1$ parts are used as the training set. The cumulative validation set error incurred by each model (i.e., incurred by the best function in each model) across the k iterations is maintained. This is known as the [Cross-Validation error](#). The model with least cross-validation error is declared as the “best” model. Now in this model, ERM using the entire dataset is performed to obtain the best candidate function that is used as the prediction function.

While the slides provide analysis of these cases, the following observations are worth to be noted:

- Practitioners from various application domains observed that k -fold cross-validation is a very good method for model selection.

- Analysis is cumbersome and is not very insightful for the Validation based algorithms.
- None of these directly consider the ideal goal of minimizing the generalization error. In the subsequent section, we present an algorithm for model selection that directly optimizes for this criteria and is guaranteed to achieve Bayes consistency.

3.2.2 Infinitely-sized Model Class: example of Bayes Consistency

Bibliography

- P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.
- B. E. Boser, I. M. Guyon, and V. N. Vapnik. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the Annual ACM Workshop on Computational Learning Theory*, pages 144–152, 1992.
- C. J .C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2), 1998.
- H. Chernoff. A Measure of Asymptotic Efficiency of Tests of a Hypothesis based on the Sum of Observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.
- A. Christmann and I. Steinwart. Universal Kernels on Non-standard Input Spaces. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.
- C. Cortes and V .N. Vapnik. Support Vector Networks. 20:273–297, 1995.
- G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley, 2 edition, 1996.
- A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 1970.
- Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004. ISSN 1532-4435.
- G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M.I. Jordan. Learning the Kernel Matrix with Semidefinite Programming. *Journal of Machine Learning Research*, 5:27–72, 2004.
- M. Ledoux and M. Talagrand. *Probability in Banach Space*. Springer-Verlag, New York, 1991.

- C. McDiarmid. On the methods of bounded differences. *Surveys in Combinatorics*, pages 148–188, 1989.
- Ron Meir and Tong Zhang. Generalization Error Bounds for Bayesian Mixture Algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.
- M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT Press, 2012.
- Walter Rudin. *Principles of mathematical analysis*. McGraw-Hill, 3rd edition, 1976.
- Bernhard Schölkopf and Alex Smola. *Learning with Kernels*. MIT press, Cambridge, 2002.
- Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.
- L. Song, A. Gretton, and K. Fukumizu. Kernel embeddings of conditional distributions. *IEEE Signal Processing Magazine*, 2013.
- Ingo Steinwart. On the Influence of the Kernel on the Consistency of Support Vector Machines. *Journal of Machine Learning Research*, 2:67–93, 2001.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.
- V. N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, Inc., 1998.