Tutorial: Classical Machine Learning Theory

Generalization, uniformly convergence, and Rademacher complexity

Chenyu Zheng[1]

[1]Gaoling School of AI
Renmin University of China

June 2023

Table of Contents

Table of Contents

1.1 What is machine learning?

Roughly speaking, learning is the process of converting experience into expertise or knowledge. The **input** to a **learning algorithm** is training data, representing experience, and the **output** is some expertise.

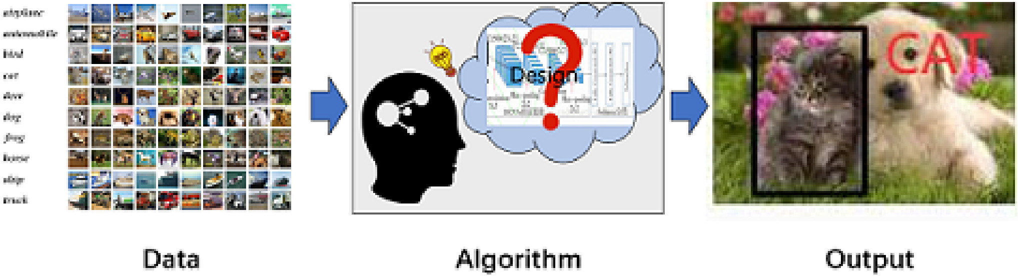

Figure: A standard learning process.

1.2 What do we need to clarify at first?

Seeking a formal-mathematical understanding of ML, we'll have to be more explicit about what we mean by each of the involved terms:

- **Data:** What is the training data our programs will access?
  - **Distribution-free**/special distribution, **i.i.d..**/non-i.i.d., ...
  - **labeled**/unlabeled, **binary**/multi-class/multi-label, **clean**/noisy, ...
- **Algorithm:** How can the process of learning be automated?
  - **Hypothesis set: finite hypotheses**/**linear models**/neural networks,
  - **Loss function**: **zero-one loss**/**convex surrogate loss**(cross-entropy loss, hinge loss) ...,
  - **Optimization: empirical risk minimization**/gradient-based/reinforcement learning, ...
- **Performance:** How can we evaluate the success of the quality of the output?
  - **Guarantee**: **Generalization bounds, sample complexity**

1.3 Mathematical definitions

We introduce some mathematical notations.

- Data:
  - Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the input space and $\mathcal{Y} = \{-1, +1\}$ be the label space.
  - We denote by $\mathcal{D}$ the population distribution over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.
  - Training set S with m $i.i.d.$ examples sampled from the distribution $\mathcal{D}$.
- Algorithm
  - Hypothesis set $\mathcal{H}$: set of functions $\mathcal{X} \to \mathcal{Y}$ (for 0-1 loss) or $\mathcal{X} \to \mathbb{R}$ (otherwise). In addition, $\mathcal{H}_{\text{all}}$ denotes the set of all functions.
  - Loss function $\ell : \mathcal{Y} \times \mathcal{Y} / \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$.
    - Zero-one loss: $\ell_{0-1}(h(\mathbf{x}), y) = \mathbb{1}_{h(\mathbf{x}) \neq y}$,
    - Logistic loss: $\ell_{\log}(h(\mathbf{x}), y) = \log_2(1 + e^{-yh(\mathbf{x})})$,
    - Empirical error $\widehat{\mathcal{R}}_{\ell,S}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h(\mathbf{x}_i), y_i)$, true error $\mathcal{R}_\ell(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h(\mathbf{x}), y)]$.
  - ERM: return hypothesis $h_S = \text{argmin}_{h\in\mathcal{H}} \widehat{\mathcal{R}}_{\ell,S}(h)$.

1.4 Roadmap of this tutorial

Given training set S, hypothesis $\mathcal{H}$, and learning algorithm $\mathcal{A}$, we can obtain a hypothesis h_S. We'd like to estimate or give some guarantees for $\mathcal{R}_\ell(h_S)$.

- (Generalization error) Estimate $\mathcal{R}_\ell(h_S)$ based on $\widehat{\mathcal{R}}_{\ell,S}(h_S)$:

$$\mathcal{R}_\ell(h_S) \leq \widehat{\mathcal{R}}_{\ell,S}(h_S) + f^1_{\mathcal{H},\mathcal{A}}(m)?$$

- (Estimation error) Distance between h_S and the optimal hypothesis in $\mathcal{H}$:

$$\mathcal{R}_\ell(h_S) \leq \inf_{h\in\mathcal{H}} \mathcal{R}_\ell(h) + f^2_{\mathcal{H},\mathcal{A}}(m)?$$

- (Excess risk) Distance between h_S and the Bayes optimal predictor:

$$\mathcal{R}_\ell(h_S) \leq \inf_{h\in\mathcal{H}_{\mathrm{all}}} \mathcal{R}_\ell(h) + f^3_{\mathcal{H},\mathcal{A}}(m)?$$

Problem: randomness

Because S is randomly sampled from distribution $\mathcal{D}$, $\mathcal{R}_\ell(h_S), \widehat{\mathcal{R}}_{\ell,S}(h_S)$ and $f^i_{\mathcal{H},\mathcal{A}}(m)$ mentioned above is not deterministic. How can we clarify this?

1.5 Concentration with high probability

To solve the problem from the randomness of these terms, we construct theoretical guarantees with high probability, but not deterministicly. Formally, with high probability $1-\delta$, we study:

- (Generalization error) Estimate $\mathcal{R}_\ell(h_S)$ based on $\widehat{\mathcal{R}}_{\ell,S}(h_S)$:

$$\mathcal{R}_\ell(h_S) \le \widehat{\mathcal{R}}_{\ell,S}(h_S) + f^1_{\mathcal{H},\mathcal{A}}(m,\delta)?$$

- (Estimation error) Distance between h_S and the optimal hypothesis in $\mathcal{H}$:

$$\mathcal{R}_\ell(h_S) \le \inf_{h\in\mathcal{H}} \mathcal{R}_\ell(h) + f^2_{\mathcal{H},\mathcal{A}}(m,\delta)?$$

- (Excess risk) Distance between h_S and the Bayes optimal predictor:

$$\mathcal{R}_\ell(h_S) \le \inf_{h\in\mathcal{H}_{\text{all}}} \mathcal{R}_\ell(h) + f^3_{\mathcal{H},\mathcal{A}}(m,\delta)?$$

A view from mathematician

These high probability bounds are closely related to "concentration inequalities"! It focuses on bounding the probability $\mathbb{P}(\left|X-\mathbb{E}[X]\right| \ge \epsilon)$. Lots of tools can be found in the literature [1, 2].

1.6 Some important concentration inequalities

Lemma 1 (Hoeffding's inequality, bound for $\sum_{i=1}^{m} X_i$, Theorem D.2, [3])

Let $X_1, \ldots, X_m$ be independent random variables with X_i taking values in $[a_i, b_i]$ for all i. Then, for any $\epsilon > 0$, the following inequalities hold for $S_m = \sum_{i=1}^{m} X_i$:

$$\mathbb{P}\left[\left|S_m - \mathbb{E}\left[S_m\right]\right| \geq \epsilon\right] \leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m}\left(b_i - a_i\right)^2}\right).$$

1.6 Some important concentration inequalities

Lemma 2 (McDiarmid's inequality, bound for $f(X_1^m)$, Theorem D.8, [3])

Let $X_1, \ldots, X_m$ be a set of m independent random variables and assume that there exist $c_1, \ldots, c_m > 0$ such that $f : X^m \to \mathbb{R}$ satisfies the following conditions:

$$\left| f\left(x_1, \ldots, x_i, \ldots, x_m\right) - f\left(x_1, \ldots, x_i', \ldots x_m\right) \right| \leq c_i$$

for all i and any points $x_1, \ldots, x_m, x_i'$. Let $f(S)$ denote $f\left(X_1, \ldots, X_m\right)$, then, for all $\epsilon > 0$, the following inequality holds:

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

1.6 Some important concentration inequalities

Lemma 3 (Maximal inequality, bound for finite $\max_j X_j$, Corollary D.11, [3])

Let $X_1 \ldots X_n$ be n real-valued random variables such that for all $j, X_j = \sum_{i=1}^m Y_{ij}$ where, for each fixed j, Y_{ij} are independent zero mean random variables taking values in $[-r_i, +r_i]$, for some $r_i > 0$. Then, the following inequality holds:

$$\mathbb{E}\left[\max_j X_j\right] \leq r\sqrt{2\log n}$$

with $r = \sqrt{\sum_{i=1}^m r_i^2}$.

Table of Contents

Uniformly convergence

First, we want to bound $\mathcal{R}_\ell(h_S) - \widehat{\mathcal{R}}_{\ell,S}(h_S)$. However, it is difficult because we do not know which hypothesis h_S is selected by the learning algorithm $\mathcal{A}$. Besides, Hoeffding's inequality can not be used directly due to the independence assumption failing to hold. Therefore, we skip this problem by giving a uniform convergence bound, that is, a bound that holds for the set of all hypotheses in $\mathcal{H}$, which a fortiori includes h_S.

Uniformly convergence

To solve the problem mentioned above, we will bound

$$\sup_{h\in\mathcal{H}}\left|\mathcal{R}_\ell(h) - \widehat{\mathcal{R}}_{\ell,S}(h)\right| \geq \left|\mathcal{R}_\ell(h_S) - \widehat{\mathcal{R}}_{\ell,S}(h_S)\right|,$$

which changes the bound from algorithm-dependent to algorithm-independent. To derive algorithm-dependent bounds for h_S, one can refer to algorithm stability [4] and information theory [5].

2.2 Finite hypothesis set

In this part, we introduce the first generalization bound in this tutorial, which focuses on the **finite hypothesis set**, that is, $|\mathcal{H}| < +\infty$. The result shows that we only need $O(\log|\mathcal{H}|)$ samples to make the generalization error small enough.

Theorem 4 (Theorem 2.13, [3])

Let $\mathcal{H}$ be a finite hypothesis set and the loss function ℓ is bounded by M. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds:

$$\forall h \in \mathcal{H}, \quad \mathcal{R}_\ell(h) \leq \widehat{\mathcal{R}}_{\ell,S}(h) + M\sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}}.$$

Problem: infinite hypothesis set

If $|\mathcal{H}| = +\infty$ (e.g., linear classifiers, neural networks), then the bound will be meaningless.

2.3 Rademacher complexity

Rademacher complexity is used to establish generalization bounds for infinite hypothesis sets. It measures the ability of the hypothesis set to capture the randomness. We note that $\mathfrak{R}_m(\mathcal{H}_{\text{all}}) = 1$ when we consider $\mathcal{G} : \mathcal{X} \mapsto \{-1, +1\}$.

Definition 5 (Rademacher complexity)

Let $\mathcal{D}$ denote the distribution according to which samples are drawn. For any integer $m \geq 1$, the Rademacher complexity of $\mathcal{G}$ is the expectation of the empirical Rademacher complexity over all samples of size m drawn according to $\mathcal{D}$:

$$\mathfrak{R}_m(\mathcal{G}) = \underset{S \sim \mathcal{D}^m}{\mathbb{E}} \left[\widehat{\mathfrak{R}}_S(\mathcal{G}) \right] = \underset{S \sim \mathcal{D}^m}{\mathbb{E}} \underset{\boldsymbol{\sigma}}{\mathbb{E}} \left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g\left(z_i\right) \right],$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)^\top$, with $\sigma_i s$ independent uniform random variables taking values in $\{-1, +1\}$. The random variables σ_i are called Rademacher variables.

Theorem 6 (Theorem 3.3, [3])

Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[0, M]$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an $i.i.d.$. sample S of size m, each of the following holds for all $g \in \mathcal{G}$:

$$\mathbb{E}[g(z)] \leq \frac{1}{m} \sum_{i=1}^{m} g\left(z_i\right) + 2\mathfrak{R}_m(\mathcal{G}) + M\sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

Relation with generalization bounds

Let $\mathcal{G} = \{(x, y) \rightarrow \ell(h(\mathbf{x}), y) : h \in \mathcal{H}\}$ with $\ell \leq M$. For any fixed $h \in \mathcal{H}$, we have

$$\mathcal{R}_\ell(h) \leq \widehat{\mathcal{R}}_{\ell,S}(h) + 2\mathfrak{R}_m(\mathcal{G}) + M\sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

Table of Contents

3.1 Bounds for the zero-one loss

In this part, we derive the generalization bound for the zero-one loss, and assume that ERM can be performed. Our task is to bound the Rademacher complexity of $\mathcal{G}$. The following lemma is used to simplify the $\mathfrak{R}_m(\mathcal{G})$ to $\mathfrak{R}_m(\mathcal{H})$.

Lemma 7 (Lemma 3.4, [3])

Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ and let $\mathcal{G}$ be the family of loss functions associated to $\mathcal{H}$ for the zero-one loss: $\mathcal{G} = \{(\mathbf{x}, y) \mapsto \mathbb{1}_{h(\mathbf{x}) \neq y} \mid h \in \mathcal{H}\}$. Then, the following relation holds between the Rademacher complexities of $\mathcal{G}$ and $\mathcal{H}$:

$$\mathfrak{R}_m(\mathcal{G}) = \frac{1}{2}\mathfrak{R}_m(\mathcal{H}).$$

3.1 Bounds for the zero-one loss

Now, we are ready to derive generalization bounds for binary classification in terms of the Rademacher complexity of the hypothesis set $\mathcal{H}$.

Theorem 8 (Theorem 3.5, [3])

Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ and ℓ be the zero-one loss. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over a sample S of size m drawn according to $\mathcal{D}$, each of the following holds for any $h \in \mathcal{H}$:

$$\mathcal{R}_{\ell_{0-1}}(h) \leq \widehat{\mathcal{R}}_{\ell_{0-1},S}(h) + \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

Problem: hardness for computing Rademacher complexity

Rademacher complexity is distribution-dependent. However, it is hard to compute.

3.1 Bounds for the zero-one loss

To bound the $\mathfrak{R}_m(\mathcal{H})$, the core property we will use is that the outputs $\{(h(\mathbf{x}_1), \ldots, h(\mathbf{x}_m)) : h \in \mathcal{H}\}$ is finite $(\mathrm{poly(m)})$ when S is fixed, though the hypothesis set $\mathcal{H}$ is infinite. Built upon this, we can define the growth function as follows.

Definition 9 (Growth function)

The growth function $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ for a hypothesis set $\mathcal{H}$ is defined by:

$$\forall m \in \mathbb{N}, \Pi_{\mathcal{H}}(m) = \max_{\{x_1,\ldots,x_m\} \subseteq \mathcal{X}} \left| \left\{ \left(h\left(\mathbf{x}_1\right), \ldots, h\left(\mathbf{x}_m\right)\right) : h \in \mathcal{H} \right\} \right| \leq 2^m.$$

Distribution-independent property of the growth function

Unlike the Rademacher complexity, this measure does not depend on the distribution $\mathcal{D}$ (uniformly with S), it is purely combinatorial, which is easier to compute or estimate.

3.1 Bounds for the zero-one loss

Then we can directly bound $\mathfrak{R}_m(\mathcal{H})$ by using Lemma 3.

Corollary 10

Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ and ℓ be the zero-one loss. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over a sample S of size m drawn according to $\mathcal{D}$, each of the following holds for any $h \in \mathcal{H}$:

$$\mathcal{R}_{\ell_{0-1}}(h) \leq \widehat{\mathcal{R}}_{\ell_{0-1},S}(h) + \sqrt{\frac{2\log \Pi_{\mathcal{H}}(m)}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

The scale of $\Pi_{\mathcal{H}}(m)$

If we naively let $\Pi_{\mathcal{H}}(m) = 2^m$, the bound is meaningless. Thus, we hope that $\Pi_{\mathcal{H}}(m) = \mathrm{poly}(m)$ to give a good guarantee.

3.1 Bounds for the zero-one loss

Definition 11 (VC-dimension)

The VC-dimension of a hypothesis set $\mathcal{H}$ is the size of the largest set that can be shattered by $\mathcal{H}$: $\mathrm{VCdim}(\mathcal{H}) = \max\left\{m : \Pi_{\mathcal{H}}(m) = 2^m\right\}$

Lemma 12 (Sauer's lemma)

Let $\mathcal{H}$ be a hypothesis set with $\mathrm{VCdim}(\mathcal{H}) = d < +\infty$, then it holds that:

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \overset{(m \geq d)}{\leq} \left(\frac{em}{d}\right)^d = O(m^d).\,(\mathrm{poly}(m))$$

Lemma 13 (VC dimension of linear models)

Let $\mathcal{H} = \left\{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle + b \mid \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\right\}$ be a linear hypothesis set in $\mathbb{R}^d$, then $\mathrm{VCdim}(\mathcal{H}) = d + 1$.

3.1 Bounds for the zero-one loss

Theorem 14 (Generalization bound w.r.t VC-dimension)

Let $\mathcal{H}$ be a family of functions taking values in $\{-1,+1\}$, $\mathrm{VCdim}(\mathcal{H}) = d < +\infty$, and ℓ be the zero-one loss. Then, for any $\delta > 0$, with probability at least $1-\delta$ over a sample S of size m drawn according to $\mathcal{D}$, the following holds for any $h \in \mathcal{H}$:

$$\mathcal{R}_{\ell_{0-1}}(h) \leq \widehat{\mathcal{R}}_{\ell_{0-1},S}(h) + \sqrt{\frac{2d}{m}\log\frac{em}{d}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

Limitations of the VC dimension

The bounds for the VC dimension always **depend on the dimension or the number of parameters**. This will not be a good bound for high(infinite)-dimensional models (e.g., kernel methods). Besides, it does not consider other properties of data (e.g., norm).

3.2 Bounds for the surrogate losses

In practice, directly optimizing the zero-one loss is NP-hard. We usually make use of some surrogate losses (ideally, convex), which are easy to optimize.

- logistic regression: logistic (cross-entropy) loss $\ell_{\log}(h(\mathbf{x}), y) = \log_2(1 + e^{-yh(\mathbf{x})})$,
- support vector machine: hinge loss $\ell_{\text{hinge}}(h(\mathbf{x}), y) = \max(0, 1 - yh(\mathbf{x}))$,
- AdaBoost: exponential loss $\ell_{\exp}(h(\mathbf{x}), y) = \exp(-yh(\mathbf{x}))$.

Relation between the zero-one loss and above surrogate losses

The above surrogate losses upper bound the zero-one loss, which implies that we can bound the true error w.r.t zero-one loss by the empirical error w.r.t. surrogate losses.

$$\mathcal{R}_{\ell_{0-1}}(h) \leq \mathcal{R}_{\ell_{\text{sur}}}(h) \leq \widehat{\mathcal{R}}_{\ell_{\text{sur}},S}(h) + 2\mathfrak{R}_m(\mathcal{G}) + M\sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

3.2 Bounds for the surrogate losses

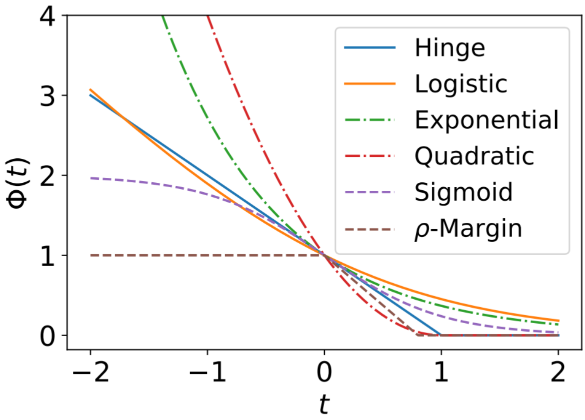

Figure: Visualization of the loss functions. t means the margin $yh(\mathbf{x})$.

3.2 Bounds for the surrogate losses

Similarly to the case with the zero-one loss, there exists a relationship between $\mathfrak{R}_m(\mathcal{G})$ and $\mathfrak{R}_m(\mathcal{H})$ when the surrogate loss is Lipschitz.

Lemma 15 (Talagrand's lemma)

Let $\phi : \mathbb{R} \to \mathbb{R}$ be a κ-Lipschitz function. Then

$$\mathfrak{R}_m(\phi \circ \mathcal{H}) \leq \kappa \mathfrak{R}_m(\mathcal{H}),$$

where $\phi \circ \mathcal{H} = \{z \mapsto \phi(h(z)) : h \in \mathcal{H}\}$.

Lipschitzness of the above surrogate losses

Logistic loss and hinge loss are Lipschitz functions w.r.t. $yh(\mathbf{x})$ (margin). Besides, the exponential loss is also Lipschitz when the $yh(\mathbf{x})$ is bounded.

Theorem 16 (Rademacher complexity of linear hypotheses with bounded ℓ_2 norm)

Let $\mathcal{H} = \left\{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle + b \mid \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_2 \leq W\right\}$ *for some constant* $W > 0$, *and* $\widetilde{\mathcal{H}} = \{(\mathbf{x}, y) \mapsto yh(\mathbf{x}) \mid h \in \mathcal{H}\}$. *Moreover, assume that* $\|\mathbf{x}\|_2 \leq C$ *(or* $\mathbb{E}\left[\|\mathbf{x}\|_2^2\right] \leq C^2$*), where* $C > 0$ *is a constant. Then*

$$\widehat{\mathfrak{R}}_S(\widetilde{\mathcal{H}}) = \widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \frac{W}{m}\sqrt{\sum_{i=1}^{m} \|\mathbf{x}_i\|_2^2},$$

and

$$\mathfrak{R}_m(\widetilde{\mathcal{H}}) = \mathfrak{R}_m(\mathcal{H}) \leq \frac{WC}{\sqrt{m}}.$$

Corollary 17 (Generalization bound for the linear hypothesis set)

Let $\mathcal{H} = \left\{ \mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle + b \mid \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_2 \le W, b \in \mathbb{R}, |b| \le B \right\}$ *for some constant* $W > 0$. *Assume that* $\|\mathbf{x}\|_2 \le C$ *(or* $\mathbb{E}\left[\|\mathbf{x}\|_2^2\right] \le C^2$*), where* C *is a positive constant. Let surrogate loss* ℓ_{sur} *be a* κ*-Lipschitz function w.r.t.* $yh(\mathbf{x})$ *and be bounded by* M*. Then*

$$\mathcal{R}_{\ell_{0-1}}(h) \le \mathcal{R}_{\ell_{\mathrm{sur}}}(h) \le \widehat{\mathcal{R}}_{\ell_{\mathrm{sur}}, S}(h) + 2\kappa \frac{WC}{\sqrt{n}} + M\sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

Comparison to the VC-dimension bounds

This bound is better since it does not have as strong of dependence on the dimension d (though B, C could dependent on d). Besides, it accounts for the norms of the model parameters and the data, which inspires us to use weight decay in practice.

3.3 Covering and packing

In terms of linear hypothesis class, we can use cauchy-schwarz inequality to bound its Rademacher complexity. However, this trick is not generally suitable, for example, deep neural nets. We need a general technique to bound the Rademacher complexity w.r.t surrogate losses.

Problem: infinite output space

The core difficulty is that the output space $\{(h(\mathbf{x}_1), \ldots, h(\mathbf{x}_m)) : h \in \mathcal{H}\}$ is infinite. Recall that when we discussed the zero-one loss, its size is $\mathrm{poly(m)}$ and then we use the Lemma 3 to obtain a bound.

Solution: discretization

We can find finite balls to cover and approximate the output space, then we use Lemma 3 on the finite selected balls. This is closely related to the covering and packing technique.

3.3 Covering and packing

Definition 18 (Covering number)

A ϵ-cover of a set $\mathcal{G}$ with respect to a metric ρ is a set $\{\theta_1, \ldots, \theta_N\} \subseteq \mathcal{G}$ such that for each $\theta \in \mathcal{G}$, there exists some $i \in \{1, \ldots, N\}$ such that $\rho(\theta, \theta_i) \leq \epsilon$. The ϵ-covering number $\mathcal{N}(\epsilon; \mathcal{G}, \rho)$ is the cardinality of the smallest ϵ-cover.

Definition 19 (Packing number)

A ϵ-packing of a set $\mathcal{G}$ with respect to a metric ρ is a set $\{\theta_1, \ldots, \theta_N\} \subseteq \mathcal{G}$ such that $\rho(\theta_i, \theta_j) > \epsilon$ for all distinct $i, j \in \{1, 2, \ldots, M\}$. The ϵ-packing number $\mathcal{M}(\epsilon; \mathcal{G}, \rho)$ is the cardinality of the largest ϵ-packing.

Lemma 20 (Relation between covering and packing number)

$$\mathcal{M}(2\epsilon; \mathcal{G}, \rho) \leq \mathcal{N}(\epsilon; \mathcal{G}, \rho) \leq \mathcal{M}(\epsilon; \mathcal{G}, \rho).$$

3.3 Covering and packing

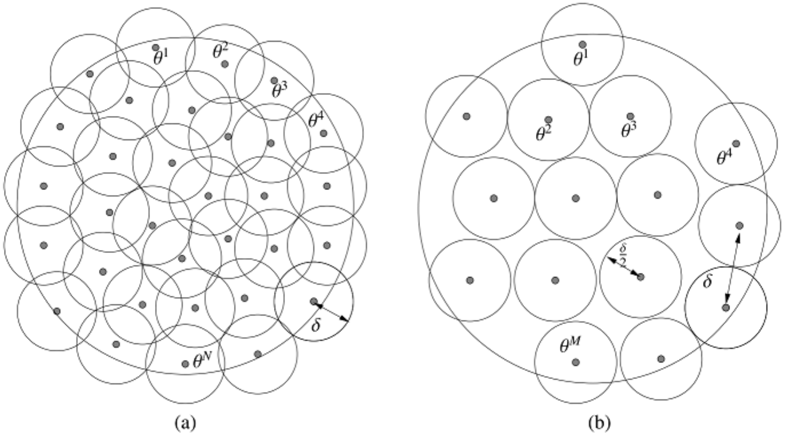

Figure: Illustration of packing and covering sets.

3.3 Covering and packing

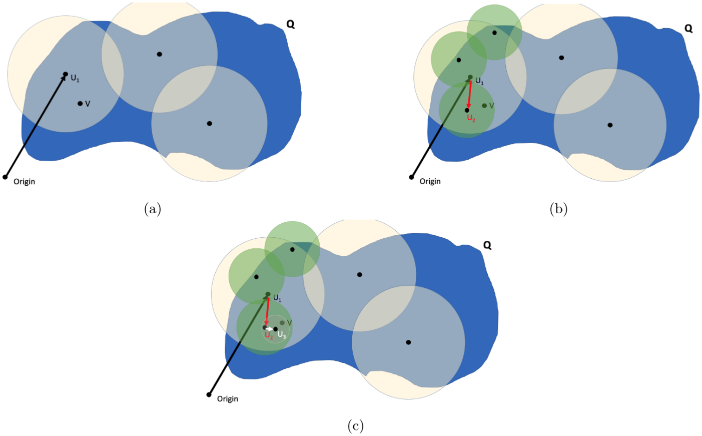

Figure: Visualization of our proof idea.

Theorem 21 (One-step discretization bound)

Let $\mathcal{H}$ be a family of functions $\mathcal{X} \mapsto [-M, M]$. Then

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \inf_{\epsilon > 0} \left(\epsilon + M \sqrt{\frac{2 \log \mathcal{N}\left(\epsilon; \mathcal{H}, L_2\left(P_n\right)\right)}{m}} \right),$$

where $L_2\left(P_n\right)\left(f, f'\right) = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(f(\mathbf{x}_i) - f'(\mathbf{x}_i)\right)^2}$. The ϵ term can be thought of as the discretization error, while the second term is the Rademacher complexity of the finite ϵ-cover.

3.3 Covering and packing

Theorem 22 (Dudley's entropy integral bound)

If $\mathcal{H}$ is a function class from $\mathcal{X} \mapsto \mathbb{R}$, then

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq 12 \int_0^\infty \sqrt{\frac{\log \mathcal{N}\left(\epsilon; \mathcal{H}, L_2\left(P_n\right)\right)}{m}} d\epsilon$$

or more generally,

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \inf_{\alpha \geq 0} \left(4\alpha + 12 \int_\alpha^\infty \sqrt{\frac{\log \mathcal{N}\left(\epsilon; \mathcal{H}, L_2\left(P_n\right)\right)}{m}} d\epsilon \right).$$

Remark

Note that unlike in Theorem 21, we do not require $h \in \mathcal{H}$ to be bounded. The remaining task is to bound the covering number $\mathcal{N}\left(\epsilon; \mathcal{H}, L_2\left(P_n\right)\right)$.

3.3 Covering and packing

We still consider the **linear hypothesis class**. We first bound the covering number of the parameter space, then obtain the covering number of the output space.

Theorem 23 (Covering number of the linear parameter space)

Let $\mathcal{W} = \left\{(\mathbf{w}, b) : \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_2 \leq W, b \in \mathbb{R}, |b| \leq B\right\}$. *Assume that* $\|x\|_2 \leq C$, *where* C *is a positive constant. Then for any* $\epsilon > 0$, *the covering number of* $\mathcal{W}$ *is bounded by*

$$\log \mathcal{N}\left(\epsilon; \mathcal{W}, \|\cdot\|_2\right) \leq d \log \left(1 + \frac{2(W+B)}{\epsilon}\right),$$

and the covering number of $\mathcal{H}$ *is bounded by*

$$\log \mathcal{N}\left(\epsilon; \mathcal{H}, L_2\left(P_n\right)\right) \leq d \log \left(1 + \frac{2(W+B)(C+1)}{\epsilon}\right).$$

3.3 Covering and packing

We conclude the Rademacher complexity of the linear hypothesis set by using the covering number as follows.

Theorem 24 (Rademacher commplexity of linear hypothesis set)

Let $\mathcal{W} = \left\{(\mathbf{w}, b) : \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_2 \leq W, b \in \mathbb{R}, |b| \leq B\right\}$*, and* $\mathcal{H} = \{\langle \mathbf{w}, \mathbf{x}\rangle + b \mid (\mathbf{w}, b) \in \mathcal{W}\}$*. Assume that* $\|x\|_2 \leq C$*, where* C *is a positive constant. Then*

$$\mathfrak{R}_m(\mathcal{H}) \leq 24\left((W + B)(C + 1)\right)\sqrt{\frac{d}{m}}.$$

Remark

This is better than the VC-dimension bound.

Table of Contents

4.1 Bounds for the estimation error

Recall that the second goal is to bound $\mathcal{R}_\ell(h_S) - \inf_{h\in\mathcal{H}} \widehat{\mathcal{R}}_\ell(h)$, which is called as **estimation error**. Let $h^* = \mathrm{argmin}_{h\in\mathcal{H}}\mathcal{R}_\ell(h)$. It can be decomposed as the following.

$$\underbrace{\mathcal{R}_\ell(h_S) - \mathcal{R}_\ell(h^*)}_{\text{estimation error}} = \underbrace{\mathcal{R}_\ell(h_S) - \widehat{\mathcal{R}}_{\ell,S}(h_S)}_{\text{generalization error}} + \widehat{\mathcal{R}}_{\ell,S}(h_S) - \widehat{\mathcal{R}}_{\ell,S}(h^*) + \widehat{\mathcal{R}}_{\ell,S}(h^*) - \mathcal{R}_\ell(h^*)$$

$$\leq \underbrace{\mathcal{R}_\ell(h_S) - \widehat{\mathcal{R}}_{\ell,S}(h_S)}_{\text{generalization error}} + \widehat{\mathcal{R}}_{\ell,S}(h^*) - \mathcal{R}_\ell(h^*)$$

Bound for $\widehat{\mathcal{R}}_{\ell,S}(h^*) - \mathcal{R}_\ell(h^*)$

The remaining task is to bound $\widehat{\mathcal{R}}_{\ell,S}(h^*) - \mathcal{R}_\ell(h^*)$, which can be realized by directly using the Hoeffding's inequality (Lemma 1).

4.1 Bounds for the estimation error

Theorem 25 (Estimation error bound for the zero-one loss)

Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$, $\mathrm{VCdim}(\mathcal{H}) = d < +\infty$, and ℓ be the zero-one loss. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over a sample S of size m drawn according to $\mathcal{D}$, the following holds for any $h \in \mathcal{H}$:

$$\mathcal{R}_{\ell_{0-1}}(h) \leq \mathcal{R}_{\ell_{0-1}}(h^*_{0-1}) + \sqrt{\frac{2d}{m} \log \frac{em}{d}} + 2\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

Limitation

Performing ERM w.r.t. the zero-one loss is not practical.

4.1 Bounds for the estimation error

Theorem 26 (Estimation error bound for the surrogate losses)

Let $\mathcal{H} = \left\{ x \mapsto \langle w, x \rangle \mid w \in \mathbb{R}^d, \|w\|_2 \leq W \right\}$ *for some constant* $W > 0$*. Assume that* $\mathbb{E}\left[\|x\|_2^2\right] \leq C^2$*, where* C *is a positive constant. Let surrogate loss* ℓ_{sur} *be a* κ*-Lipschitz function w.r.t.* $yh(\mathbf{x})$ *and be bounded by* M*. Then*

$$\mathcal{R}_{\ell_{0-1}}(h) \leq \mathcal{R}_{\ell_{\mathrm{sur}}}(h) \leq \mathcal{R}_{\ell_{\mathrm{sur}},S}(h^*_{\mathrm{sur}}) + 2\kappa \frac{WC}{\sqrt{n}} + 2M\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

Limitation

In practical classification tasks, we care more about $\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}_{\ell_{0-1}}(h^*_{0-1})$ than $\mathcal{R}_{\ell_{\mathrm{sur}}}(h) - \mathcal{R}_{\ell_{\mathrm{sur}}}(h^*_{\mathrm{sur}})$. However, we can not obtain the information about the former.

4.1 Bounds for the estimation error

$\mathcal{H}$-consistency bound tries to build the quantitative relationship between $\mathcal{R}_{\ell_{0-1}}(h) - \mathcal{R}_{\ell_{0-1}}(h^*_{0-1})$ and $\mathcal{R}_{\ell_{\text{sur}}}(h) - \mathcal{R}_{\ell_{\text{sur}}}(h^*_{\text{sur}})$.

Definition 27

$\mathcal{H}$-consistency bound is in the following form that holds for all $h \in \mathcal{H}$ and some non-decreasing function $f : \mathbb{R}_+ \to \mathbb{R}_+$:

$$\mathcal{R}_{\ell_2}(h) - \mathcal{R}_{\ell_2}(h^*_{\ell_2}) \leq f\left(\mathcal{R}_{\ell_1}(h) - \mathcal{R}_{\ell_1}(h^*_{\ell_1})\right).$$

If $f(0) = 0$, then we call the loss ℓ is $\mathcal{H}$-consistent, which is important in reality.

Existing works

[6] constructs a tight binary $\mathcal{H}$-consistency framework for different hypotheses and losses. [7, 8] further extend it to the multiclass case in different ways.

4.2 Discussion about the excess risk

Our final goal is to discuss the excess risk $\mathcal{R}_\ell(h_S) - \inf_{h\in\mathcal{H}_{\text{all}}}\mathcal{R}_\ell(h)$. We define $h^* = \text{argmin}_{h\in\mathcal{H}}\mathcal{R}_\ell(h)$ and $h_{\text{Bayes}} = \text{argmin}_{h\in\mathcal{H}_{\text{all}}}\mathcal{R}_\ell(h)$. We can decompose the excess risk as:

$$\underbrace{\mathcal{R}_\ell(h_S) - \mathcal{R}_\ell(h_{\text{Bayes}})}_{\text{excess risk}} = \underbrace{\mathcal{R}_\ell(h_S) - \mathcal{R}_\ell(h^*)}_{\text{estimation error}} + \underbrace{\mathcal{R}_\ell(h^*) - \mathcal{R}_\ell(h_{\text{Bayes}})}_{\text{approximate error}}.$$

Bias-complexity trade-off

- Estimation error depends on the training set size and the complexity of the hypothesis set. It increases as the hypothesis set becomes more complex (overfitting).
- Approximate error is determined by the hypothesis class chosen. Reducing the hypothesis class can increase the approximation error (underfitting).
- We need inductive biases to select a good hypothesis class (MLP vs. CNN).

4.2 Discussion about the excess risk

By setting $\mathcal{H} = \mathcal{H}_{\text{all}}$, we can obtain the definition of the Bayes consistency.

Definition 28

Bayes consistency bound is in the following form that holds for all measurable h and some non-decreasing function $f : \mathbb{R}_+ \to \mathbb{R}_+$:

$$\mathcal{R}_{\ell_2}(h) - \mathcal{R}_{\ell_2}(h_{\ell_2,\text{Bayes}}) \leq f\left(\mathcal{R}_{\ell_1}(h) - \mathcal{R}_{\ell_1}(h_{\ell_1,\text{Bayes}})\right).$$

If $f(0) = 0$, then we say the loss ℓ is Bayes-consistent, which is important in reality.

Existing works

[9, 10] analyze the relationship between the excess risk of zero-one loss and that of a surrogate loss, and prove that lots of convex surrogate losses are Bayes-consistent.

References I

- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
- Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press, 2019.
- Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.
- Olivier Bousquet and André Elisseeff. "Stability and Generalization". In: *J. Mach. Learn. Res.* 2 (2002), pp. 499–526.
- Aolin Xu and Maxim Raginsky. "Information-theoretic analysis of generalization capability of learning algorithms". In: *Advances in Neural Information Processing Systems*. 2017, pp. 2524–2533.
- Pranjal Awasthi et al. "$\mathcal{H}$-Consistency Bounds for Surrogate Loss Minimizers". In: *International Conference on Machine Learning*. Vol. 162. 2022, pp. 1117–1174.

References II

- Pranjal Awasthi et al. "Multi-Class $\mathcal{H}$-Consistency Bounds". In: *Advances in Neural Information Processing Systems*. 2022.
- Chenyu Zheng et al. "Revisiting Discriminative vs. Generative Classifiers: Theory and Implications". In: *CoRR* abs/2302.02334 (2023).
- Tong Zhang. "Statistical behavior and consistency of classification methods based on convex risk minimization". In: *The Annals of Statistics* 32.1 (2004), pp. 56–85.
- Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. "Convexity, classification, and risk bounds". In: *Journal of the American Statistical Association* 101.473 (2006), pp. 138–156.