CUHK STAT6050: Statistical Learning Theory Spring 2026

Lecture 10: Fisher consistency

Binary classification, surrogate loss and excess risk bounds

Lecturer: Ben Dai

"*There is Nothing More Practical Than A Good Theory.*" — Kurt Lewin

1 History of Fisher Consistency

The concept of consistency originates from [Fisher, 1922], which is now regarded as the first principle of modern statistics. In fact, the development of this now widely accepted concept was not straightforward; in this section, we review the origins and evolution of consistency.

First, let's quote the following statements to illustrate Fisher's original conception of consistency.

> Section 2 in [Fisher, 1925]:
>
> *A statistic is said to be a consistent estimate of any parameter, if when calculated from an indefinitely large sample it tends to be accurately equal to that parameter.*

This initial concept eventually evolved into the definition of (probability) consistency.

Definition 1.1 (PC; Probability Consistency)**.** Given a dataset $(X_1, \cdots, X_n)$, suppose $\theta_n = T(X_1, \cdots, X_n)$ is an estimator of a parameter θ^*, then θ_n is (probability) consistent, if for every $\varepsilon > 0$,

$$\mathbb{P}(|\theta_n - \theta^*| > \varepsilon) \to 0, \text{ as } n \to \infty.$$

Furthermore, PC is also commonly referred to as convergence in probability to the true parameter, denoted by $\theta_n \xrightarrow{\mathbb{P}} \theta^*$.

PC appears to be intuitive and impeccable, as it represents our most fundamental requirement for an estimator. However, establishing a widely accepted definition of PC was not a straightforward process in the history of statistics. Now, upon recalling Fisher's original definition of consistency in [Fisher, 1922], we find that it substantially diverges from the modern definition of PC presented in Definition 1.1.

The following quote illustrates the consistency definition in [Fisher, 1922].

> Definition section in [Fisher, 1922]:
>
> *Consistency.—A statistic satisfies the criterion of consistency, if, when calculated from the whole population, it is equal to the required parameter.*

This definition appears somewhat ambiguous due to the unclear implementation of the phrase "when calculated from the whole population." To clarify, it is necessary to adopt the Fisherian perspective, which typically regards an estimator as a functional of the empirical cumulative distribution function (cdf; $\widehat{F}_n$), denoted as $\theta_n = T(\widehat{F}_n)$. By adopting this perspective, we can find a clue to interpreting the phrase, namely by replacing $\widehat{F}_n$ with F, and then verifying whether $T(F)$ equals the true parameter θ^*. This initiative eventually evolved into Fisher Consistency (FC). Subsequent discussions in the literature, including contributions from [Kallianpur, 1955, Rao, 1962, Savage, 1976, Geisser, 1980, Fienberg, 2012], led to the formal definition of FC.

Definition 1.2 (FC; Fisher Consistency)**.** Given an empirical cdf $\widehat{F}_n$, suppose $\theta_n = T(\widehat{F}_n)$ is an estimator of θ^*, then θ_n is Fisher consistent if $T(F) = \theta^*$.

The question naturally arises as to whether FC is defined to match PC. To understand the "legitimacy" of this definition, we revisit Fisher's initial intuition as the following equation: (i) Fisher seemingly assumed that $T(\widehat{F}_n)$ would eventually converge to $T(F)$; (ii) therefore, to obtain PC, we only need to require that $T(F) = \theta^*$, which, in turn, induced the definition of FC.

$$\theta_n = \overbrace{T(\widehat{F}_n) \xrightarrow{\mathbb{P}} T(F)}^{\text{Fisher's assumption}} \underbrace{= \theta^*}_{\text{FC}}. \tag{1}$$

Thus, we reasonably hypothesize that Fisher was also pursuing PC when defining FC, but decomposed PC as in (1), thereby deriving a simplified condition, FC. However, this approach relies on Fisher's default assumption, $T(\widehat{F}_n) \xrightarrow{\mathbb{P}} T(F)$, which *does not always hold*.

Specifically, according to the Glivenko-Cantelli theorem [Glivenko, 1933, Cantelli, 1933], $\widehat{F}_n \to F$ almost surely. Thus, the assumption is analogous to the conclusion of the continuous mapping theorem for $\widehat{F}_n$, where we typically require certain "continuity" conditions on $T(\cdot)$. Therefore, when rigorously examining the relationship between FC and PC through mathematical definitions, the conclusions can be rather uninteresting. There are always counterexamples where FC does not imply PC, or vice versa.

Remark 1.3. We provide some evidence from the literature to corroborate Fisher's initial idea regarding the definition of FC. Page 333 in [Kallianpur, 1955]: "*It is, thus, clear by consistency Fisher has in mind both the properties of the statistic tending to a limit in probability and limiting value being attained by the statistic at the expected value of the frequencies...*"

However, more practically, we are more interested in the positive scenario. Specifically, when $T(\cdot)$ is continuous (a property satisfied by most reasonable estimators), FC and PC are naturally equivalent. In this context, FC demonstrates its significant advantages, as it is much easier to verify than PC, since FC provides a *deterministic equation*.

Based on the above discussion, we can summarize the advantages and disadvantages of FC and PC as in Table 1. The primary limitation of FC lies in its restricted applicability, being solely suitable for estimators based on empirical cdfs. On the other hand, for most continuous estimators, FC can significantly simplify PC procedure by using only one single equality.

FC: Pros	FC: Cons
• The FC equality $T(F) = \theta^*$ is often easy to verify. • FC, in turn, serves as a condition for designing estimators. • The PC can be directly derived if T is continuous.	• FC was solely defined for estimators based on the empirical cdf $\widehat{F}_n$. • There exist certain T for which PC and FC are not equivalent.

PC: Pros	PC: Cons
• Truly reflects the estimator's asymptotic behavior.	• It is overly cumbersome for practical application.

Table 1: Advantages and disadvantages of Fisher Consistency (FC) and Probability Consistency (PC).

2 Risk-based Consistency in Machine Learning

In this section, we extend the concept of FC to machine learning (ML) problems. To illustrate, we give some examples and the generic setup of ML problems.

Example 2.1 (MSE regression)**.** *Consider a regression problem where we have a dataset consisting of input feature vectors* $\mathbf{X} \in \mathscr{X} \subset \mathbb{R}^d$ *and corresponding output values* $Y \in \mathbb{R}$*. The goal is to learn a regression function* f *that accurately predicts* Y *given* $\mathbf{X}$*. The performance of* f *is evaluated using the mean square error (MSE):*

$$MSE(f) = \mathbb{E}\Big(\big(Y - f(\mathbf{X})\big)^2\Big).$$

Example 2.2 (MAE regression)**.** *All settings are the same as Example 2.1, except that the evaluation metric is mean absolute error (MAE):*

$$MAE(f) = \mathbb{E}\Big|Y - f(\mathbf{X})\Big|.$$

Example 2.3 (Binary classification)**.** *Consider a binary classification problem where we have a dataset consisting of input feature vectors* $\mathbf{X} \in \mathscr{X} \subset \mathbb{R}^d$ *and corresponding binary labels* $Y \in \{-1, 1\}$*. The goal is to learn a classification function* f *that accurately predicts the binary label* Y *given* $\mathbf{X}$*. The performance of* f *is evaluated using classification error:*

$$MCE(f) = 1 - Acc(f) = \mathbb{E}\Big(\mathbf{1}\big(Yf(\mathbf{X}) \le 0\big)\Big),$$

where $\mathbf{1}(\cdot)$ *is the indicator function.*

According to Examples 2.1–2.3, when formulating a ML problem, three essential aspects must be clarified: (i) the input and output data, (ii) the predictive function f, and (iii) the evaluation metric for assessing the performance of predictions. Therefore, we can provide a generic setup of ML problems as follows:

- Data: $(\mathbf{X}, Y)$, where $\mathbf{X} \in \mathscr{X} \subset \mathbb{R}^d$ is the input feature vector, and $Y \in \mathbb{R}$ is the output.
- Aim: learn a predictive function $f : \mathscr{X} \to \mathbb{R}$, such that $f(\mathbf{X})$ accurately predicts Y in terms of the evaluation metric.
- Evaluation: the performance of f is evaluated by a *risk function*:

$$R(f) = \mathbb{E}\Big(L\big(Y, f(\mathbf{X})\big)\Big),$$

  where $L(\cdot)$ is a loss function to measure the "distance" between the true outcome Y and the predicted value $f(\mathbf{X})$.

In this setting, we highlight several key differences between machine learning problems and traditional statistical estimators. First, the parameter estimator, θ_n, is generalized to a non-parametric predictive function, $\widehat{f}_n$. Second, our focus has shifted from *parameter consistency* to performance or *risk-based consistency* in terms of the evaluation metric $R(\cdot)$, as our primary goal in ML problems is to optimize prediction performance rather than estimating a specific estimator or parameter. Thus, the ground truth parameter θ^* is now replaced by the *optimal risk* defined as:

$$R^* = \inf_f R(f),$$

where the minimization is taken over all measurable predictive functions f. Specifically, for practically characterizing convergence in ML problems, we also focus on *risk-based convergence*, analogous to weak convergence in functional analysis [Ciarlet, 2013]. Indeed, we adopt the following definition:

Definition 2.4 (Risk-based Probability Consistency)**.** Given a dataset $\mathscr{D}_n = (\mathbf{X}_i, Y_i)_{i=1}^n$, suppose $\widehat{f}_n = \mathscr{T}(\mathscr{D}_n)$ is an estimated predictive function, then $\widehat{f}_n$ is risk-based probability consistent w.r.t. the risk function $R(\cdot)$, if

$$R(\widehat{f}_n) \xrightarrow{\mathbb{P}} R^*,$$

as n approaches infinity.

Again, compared to the definition of probability consistency (Definition 1.1) in parameter estimation, the major extensions of Definition 2.4 are two-fold: (i) the shift from a parameter estimator $T(\cdot)$ to a non-parametric estimator $\mathscr{T}(\cdot)$, and (ii) the shift from parameter consistency to risk-based consistency.

We now consider extending the definition of Fisher consistency in the risk-based setting. Similarly, adopting a Fisherian perspective, suppose $\widehat{f}_n$ is a function of the empirical cdf $\widehat{F}_n$ of $(\mathbf{X},Y)$, i.e., $\widehat{f}_n = \mathscr{T}(\widehat{F}_n)$, then we get the same diagram:

$$R_n = R(\widehat{f}_n) = \overbrace{R\big(\mathscr{T}(\widehat{F}_n)\big) \xrightarrow{\mathbb{P}} R\big(\mathscr{T}(F)\big)}^{\text{FC assumption}} \underbrace{= R^*}_{\text{FC}}. \tag{2}$$

This yields the definition of Fisher consistency in the risk-based setting, which we refer to as risk-based Fisher consistency (RFC).

Definition 2.5 (Risk-based Fisher Consistency; RFC)**.** Given an empirical cdf $\widehat{F}_n$ of $(\mathbf{X},Y)$, suppose $\widehat{f}_n = \mathscr{T}(\widehat{F}_n)$ is an estimated predictive function, then $\widehat{f}_n$ is risk-based Fisher consistent w.r.t. the risk function $R(\cdot)$, if

$$R\big(\mathscr{T}(F)\big) = R^*. \tag{3}$$

Similar to the conclusion we drew in Section 1, the risk-based FC can also greatly simplify the "usage cost" of the risk-based PC; the conclusions summarized in Table 1 remain valid here as well. While in practice, the focus is on utilizing (3) to derive ML methods, which we refer to as the *RFC rule* (when there is no ambiguity, we simply call it the *FC rule*).

Remark 2.6. To bolster our confidence in applying the FC rule, one crucial question must be resolved: *whether the FC assumption holds true*. This is essentially a generalization of the Glivenko-Cantelli theorem, where $\widehat{F}_n$ uniformly converges to F. Yet when composed with $R(\cdot)$ and $\mathscr{T}(\cdot)$, can the convergence still be preserved? This question can be formulated as an *empirical process*:

$$\|\widehat{F}_n - F\|_{\mathscr{H}} \xrightarrow{\mathbb{P}} 0 \quad \text{as } n \to \infty,$$

where $\mathscr{H}$ is the functional space composed by $R(\cdot)$ and $\mathscr{T}(\cdot)$. Typically, to ensure convergence, it is necessary to impose certain restrictions on model complexity, akin to Donsker classes, which translates into constraints on $\mathscr{F}$.

In the following sections, we illustrate how the FC rule applies to binary classification.

3 Binary classification

We denote a vector of features as $\mathbf{X} \in \mathscr{X} \subset \mathbb{R}^d$, and a binary outcome (label) as $Y \in \{-1,1\}$. Our goal is to find a binary decision function $\psi : \mathscr{X} \to \{-1,1\}$ to predict a label given a new instance, and its performance is evaluated by the misclassification rate (MCR):

$$\mathbb{P}\big(Y \neq \psi(\mathbf{X})\big) = \mathbb{E}\mathbf{1}\big(Y\psi(\mathbf{X}) \leq 0\big).$$

Since ψ is a binary-valued function, to facilitate computation, we make decisions by taking the sign of a continuous function $f : \mathscr{X} \to \mathbb{R}$, that is, $\psi(\mathbf{x}) = \operatorname{sgn}(f(\mathbf{x}))$. Then, the MCR loss and its risk function can be rewritten as:

$$l(Yf(\mathbf{X})) = \mathbf{1}\big(Yf(\mathbf{X}) \leq 0\big), \quad R(f) = \mathbb{E}\big(l(Yf(\mathbf{X}))\big).$$

Remark 3.1. In binary classification, the loss function can be reduced to a univariate function of $Yf(\mathbf{X})$.

Bayes classifier

With the same idea, we first consider the Bayes classifier based on misclassification error.

Before proceeding, we assume $\mathbb{P}\big(Y=1|\mathbf{X}=\mathbf{x}\big)\neq 1/2$ for all $\mathbf{x}\in\mathscr{X}$ to simplify the notation. Note that when $\mathbb{P}\big(Y=1|\mathbf{X}=\mathbf{x}\big)=1/2$, the prediction of Y can be arbitrary.

Lemma 3.2 (Bayes classifier)**.** *f^* is a global minimizer of $R(f)$ if and only if*

$$\operatorname{sgn}\big(f^*(\mathbf{x})\big)=\operatorname{sgn}\big(\eta(\mathbf{x})-1/2\big),$$

where $\eta(\mathbf{x})=\mathbb{P}\big(Y=1|\mathbf{X}=\mathbf{x}\big)$, and $\operatorname{sgn}(u)=1$ if $u\geq 0$, -1 otherwise.

Remark 3.3 (Identifiability). In classification, we usually do not consider the "identifiability" issue, partly because the true decision function is not unique. On the other hand, we are only interested in the performance (MCR).

Remark 3.4 (Plug-in prediction). A simple method motivated by Bayes classifier is using plug-in estimator: $\widehat{\psi}(\mathbf{x})=\operatorname{sgn}(\widehat{\eta}(\mathbf{x})-1/2)$, where $\widehat{\eta}(\cdot)$ is an estimator of $\eta(\cdot)$. **PS**: The relationship/difference between classification and regression.

Empirical risk minimization

Next, given training samples $(\mathbf{x}_i,y_i)_{i=1,\cdots,n}$, we give the ERM or R-ERM on binary classification based on MCR.

$$\min_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}l(Y_if(\mathbf{x}_i))+\lambda_n\|f\|_{\mathscr{F}}^2,$$

where $\mathscr{F}$ is a candidate function class, which can be RKHS, boosting, tree methods, or neural networks. However, note that the indicator in the proposed ERM is discontinuous, which is infeasible/difficult to handle in optimization. To facilitate computation, we replace the indicator function with a surrogate loss ϕ.

$$\widehat{f}_n=\arg\min_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\phi(Y_if(\mathbf{x}_i))+\lambda_n\|f\|_{\mathscr{F}}^2, \tag{4}$$

Remark 3.5. The surrogate loss framework summarizes (and was also inspired by) a variety of classification methods, including logistic regression [Cox, 1972], AdaBoost [Freund and Schapire, 1997], and support vector machines (SVM; [Cortes and Vapnik, 1995]).

- Exponential loss: (i.e., AdaBoost)

$$\phi\big(Yf(\mathbf{X})\big)=\exp(-Yf(\mathbf{X})).$$

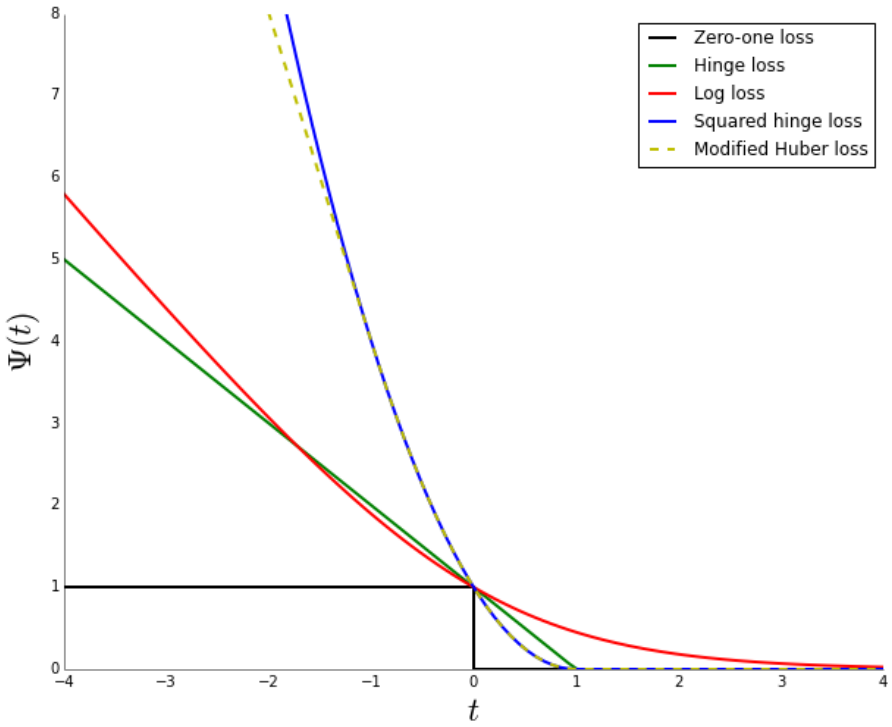

Figure 1: Plot for various surrogate losses from [Pedregosa, 2014].

- Logistic loss: (i.e., logistic regression)

$$\phi\big(Yf(\mathbf{X})\big) = \log\big(1+\exp(-Yf(\mathbf{X}))\big).$$

- Square loss:

$$\phi\big(Yf(\mathbf{X})\big) = (1-Yf(\mathbf{X}))^2.$$

- Hinge loss: (i.e., SVM)

$$\phi\big(Yf(\mathbf{X})\big) = (1-Yf(\mathbf{X}))_+.$$

In the same manner, the risk function and excess risk based on a surrogate loss ϕ are defined as:

$$R_\phi(f) = \mathbb{E}\big(\phi(Yf(\mathbf{X}))\big), \quad \mathscr{E}_\phi(f) = R_\phi(f) - R_\phi(f_\phi^*),$$

where f_ϕ^* is a minimizer of $R_\phi(f)$.

A "good" surrogate loss

Note that $\widehat{f}_n$ from (4) is no longer a minimizer of ERM based on the evaluation loss $l(\cdot)$, but rather based on a surrogate loss $\phi(\cdot)$. Therefore, the question of interest is whether we still have the "nice" asymptotics of $\mathscr{E}(\widehat{f}_n)$. We address this question in several steps.

- **Fisher consistency.** First, we consider the weakest possible condition on ϕ in population level. For every minimizer f_ϕ^* of $R_\phi(f)$, $f_\phi^*(\mathbf{x})$ should have the same sign as Bayes decision rule $\operatorname{sgn}(\eta(\mathbf{x})-1/2)$.
- **Relation between $\mathscr{E}(f)$ and $\mathscr{E}_\phi(f)$.** Since the $\widehat{f}_n$ is obtained by minimizing $\widehat{R}_{\phi,n}$, thus it is expected that $\mathscr{E}_\phi = o_P(1)$. It will be quite useful if we can obtain the relation between $\mathscr{E}(f)$ and $\mathscr{E}_\phi(f)$.
- **Convergence rates and probabilistic bounds.** The final aim is to give the convergence rate and probabilistic bound of $\mathscr{E}(\widehat{f}_n)$.

4 Fisher consistency

In this section, we give the formal definition of Fisher consistency (sometimes referred to as *classification calibration*).

Definition 4.1 (Fisher consistency [Lin, 2004])**.** A surrogate loss $\phi(\cdot)$ is Fisher consistent with respect to $l(\cdot)$ if for any $f_\phi^* \in \arg\min_f R_\phi(f)$,

$$f_\phi^*(\mathbf{X}) > 0, \text{ if } \eta(\mathbf{X}) > 1/2, \quad f_\phi^*(\mathbf{X}) < 0, \text{ if } \eta(\mathbf{X}) < 1/2, \quad \text{almost surely.}$$

Lemma 4.2. *The following surrogate losses are all Fisher consistent, and their corresponding minimizers are given as:*

- ***Logistic loss.***
$$f_\phi^*(\mathbf{x}) = \sigma^{-1}(\eta(\mathbf{x})),$$
where $\sigma(u) = 1/(1+\exp(-u))$ *is a sigmoid function.*

- ***Exponential loss.***
$$f_\phi^*(\mathbf{x}) = \big(\log(\eta(\mathbf{x})) - \log(1-\eta(\mathbf{x}))\big)/2.$$

- ***Square loss.***
$$f_\phi^*(\mathbf{x}) = 2\eta(\mathbf{x}) - 1.$$

- ***Hinge loss.***
$$f_\phi^*(\mathbf{x}) = \operatorname{sgn}(\eta(\mathbf{x}) - 1/2).$$

From Lemma 4.2, we summarize the **point-wise minimization** procedure to verify Fisher consistency of ϕ.

- **Point-wise minimization.**
$$\mathbb{E}\phi\big(Y, f(\mathbf{X})\big) = \mathbb{E}_{\mathbf{X}}\mathbb{E}\Big(\phi\big(Yf(\mathbf{X})\big)\Big|\mathbf{X}\Big) = \mathbb{E}_{\mathbf{X}}\Big(\eta(\mathbf{X})\phi\big(f(\mathbf{X})\big) + (1-\eta(\mathbf{X}))\phi\big(-f(\mathbf{X})\big)\Big),$$
it suffices to consider the point-wise minimization:
$$H(\eta) = \inf_{\alpha\in\mathbb{R}} C_\eta(\alpha) = \inf_{\alpha\in\mathbb{R}} \eta\phi\big(\alpha\big) + (1-\eta)\phi(-\alpha),$$
where $H(\eta)$ is the optimal conditional ϕ-risk. Then, for any $\mathbf{x}\in\mathscr{X}$,
$$f_\phi^*(\mathbf{x}) = \arg\min_\alpha C_\eta(\alpha).$$

Theorem 4.3 (**Convex** Fisher consistent surrogate loss [Bartlett et al., 2006])**.** *Let* $\phi(\cdot)$ *be convex. Then* ϕ *is Fisher consistent if and only if it is differentiable at 0 and* $\phi'(0) < 0$.

5 Relation between $\mathscr{E}$ and $\mathscr{E}_\phi$

Next, we investigate the relation between $\mathscr{E}$ and $\mathscr{E}_\phi$. First, we present some useful facts about $\mathscr{E}$.

Lemma 5.1. *For* $R(f)$ *and* $\mathscr{E}(f)$ *defined in Section 3, we have*
$$R^* = R(f^*) = \mathbb{E}\Big(\min(\eta(\mathbf{X}), 1-\eta(\mathbf{X}))\Big) = \big|\eta(\mathbf{X}) - 1/2\big| + 1/2,$$
$$\mathscr{E}(f) = R(f) - R(f^*) = \mathbb{E}\Big(\mathbf{1}\big\{\operatorname{sgn}\big(f(\mathbf{X})\big) \neq \operatorname{sgn}\big(f^*(\mathbf{X})\big)\big\}\Big|2\eta(\mathbf{X}) - 1\Big|\Big).$$

To investigate the relation between $\mathscr{E}$ and $\mathscr{E}_\phi$, we give the definition of the optimal disagreement risk $H^-(\eta)$ [Zhang, 2004]:
$$H^-(\eta) = \inf_{\alpha:\alpha(2\eta-1)\leq 0} C_\eta(\alpha).$$

Remark 5.2. Some important properties of $H(\eta)$ and $H^-(1-\eta)$.

- Fisher consistency is equivalent to $H^-(\eta) > H(\eta)$.
- $H(\eta) = H(1-\eta)$ and $H^-(\eta) = H^-(1-\eta)$.

Theorem 5.3 ([Zhang, 2004, Bartlett et al., 2006]). *For any nonnegative loss function ϕ, any measurable function $f : \mathscr{X} \to \mathbb{R}$ and any probability distribution on $\mathscr{X} \times \{-1,+1\}$,*

$$\psi\big(R(f) - R^*\big) \leq R_\phi(f) - R_\phi^*,$$

where $\psi : [0,1] \to [0,\infty)$ is defined as the Fenchel-Legendre biconjugate of $\widetilde{\psi}$, denoted as $\psi = \widetilde{\psi}$ and

$$\widetilde{\psi}(\theta) = H^-(\frac{1+\theta}{2}) - H(\frac{1+\theta}{2}).$$

Remark 5.4. If ψ is invertible on $[0,1]$, then

$$\mathscr{E}(f) = R(f) - R^* \leq \psi^{-1}\big(R_\phi(f) - R_\phi^*\big) = \psi^{-1}\big(\mathscr{E}_\phi(f)\big),$$

which tells the relation between $\mathscr{E}$ and $\mathscr{E}_\phi$. More importantly, it yields that good performance in $\phi(\cdot)$ yields a good performance on $l(\cdot)$.

Theorem 5.5 (Excess risk bound for convex surrogate loss [Bartlett et al., 2006]). *If ϕ is convex and Fisher consistent, then*

$$\psi(u) = \phi(0) - H(\frac{1+u}{2}).$$

6 Examples

[Bartlett et al., 2006] illustrates some examples of surrogate losses in classification.

- Exponential loss: $\phi(u) = \exp(-u)$.
  - Fisher consistency (Theorem 2.3).
  - Excess risk bound (Theorem 3.5).
    - $H(\eta) = 2\sqrt{\eta(1-\eta)}$.
    - $\psi(u) = 1 - \sqrt{1-u^2}$, for $u \in [0,1]$.
    - $\psi^{-1}(u) = \sqrt{u(2-u)}$, for $u \in [0,1]$.
    - Excess risk bound:
    $$\mathscr{E}(f) \leq \sqrt{2\mathscr{E}_\phi(f)}.$$
- Hinge loss: $\phi(u) = (1-u)_+$.
  - Fisher consistency (Theorem 2.3).

- Excess risk bound (Theorem 3.5).
  - $H(\eta) = 2\min\{\eta, 1-\eta\}$.
  - $\psi(u) = |u| = u$, for $u \in [0,1]$.
  - $\psi^{-1}(u) = u$, for $u \in [0,1]$.
  - Excess risk bound:

$$\mathscr{E}(f) \leq \mathscr{E}_\phi(f).$$

Remark 6.1. We can show that: (i) for the squared loss:

$$\mathscr{E}(f) \leq \sqrt{\mathscr{E}_\phi(f)};$$

(ii) for the logistic loss:

$$\mathscr{E}(f) \leq \sqrt{2\mathscr{E}_\phi(f)}.$$

7 A loose excess risk upper bound for kernel SVM

According to the results in the preceding sections, suppose we find an increasing function $\psi : [0,1] \to \mathbb{R}^+$, such that for any $f \in \mathscr{F}$:

$$\psi\big(\mathscr{E}(f)\big) \leq \mathscr{E}_\phi(f),$$

then, for $0 \leq \varepsilon_n \leq 1$, we have

$$\mathbb{P}\big(\mathscr{E}(\widehat{f}_n) \geq \varepsilon_n\big) = \mathbb{P}\Big(\psi\big(\mathscr{E}(\widehat{f}_n)\big) \geq \psi(\varepsilon_n)\Big) \leq \mathbb{P}\Big(\mathscr{E}_\phi(\widehat{f}_n) \geq \psi(\varepsilon_n)\Big). \tag{5}$$

Let $t_n = \psi(\varepsilon_n)$, it suffices to investigate the asymptotics of $\mathscr{E}_\phi(\widehat{f}_n)$.

R-ERM for kernel SVMs

For illustration, let's consider the kernel-based SVM, that is, SVM in RKHS:

$$\widehat{f}_n = \arg\min_{f \in \mathscr{H}_K} \frac{1}{n}\sum_{i=1}^n \big(1 - Y_i f(\mathbf{X}_i)\big)_+ + \lambda_n \|f\|^2_{\mathscr{H}_K}$$

Remark 7.1. For kernel SVM, we still have Representer theorem.

The reader may check Sections 4.5, 12.1-12.3 in [Hastie et al., 2001] to see the motivation and detailed computation of SVMs.

Remark 7.2. Truncation of $\widehat{f}_n(\mathbf{x})$ to $[-1,1]$ always yields a lower (or equal) loss.

On this ground, we assume $\|\widehat{f}_n\|_\infty$ is bounded by 1, otherwise we could consider the asymptotics of truncated estimator. Recall the decomposition:

$$\mathscr{E}_\phi(\widehat{f}_n) = R_\phi(\widehat{f}_n) - R_\phi(f^*) \leq 2\sup_{f \in \mathscr{F}} |\widehat{R}_{\phi,n}(f) - R_\phi(f)| + \mathbf{Approx}_\phi(\lambda_n),$$

where

$$\mathbf{Approx}_\phi(\lambda_n) = \inf_{f \in \mathscr{H}_n} R_\phi(f) - R_\phi(f^*) + \lambda_n \|f\|^2_{\mathscr{H}_n}.$$

Approximation error

In the regression case, we know that the convergence rate of the approximation error is related to the "smoothness" of the Bayes decision function. In the same manner, we have the following definition of "smoothness" in classification, which is the so-called "geometric-noise" assumption.

Theorem 7.3 (Theorem 2.7 in [Steinwart and Scovel, 2007]). *Let $\mathscr{X} \subset \mathbb{R}^d$ be a compact domain, K be a Gaussian kernel with a hyperparameter σ_n, and the distribution of $(\mathbf{X}, \mathbf{Y})$ satisfies the geometric-noise assumption (Definition 2.3 in [Steinwart and Scovel, 2007]) with geometric noise exponent $0 < \alpha < \infty$. Then, there exists a constant A_0, such that*

$$\mathbf{Approx}_\phi(\lambda_n) \leq A_0 \lambda_n^{\frac{\alpha}{\alpha+1}},$$

provided that $\sigma_n = \lambda_n^{-\frac{1}{d(\alpha+1)}}$.

Estimation error

For any f and f' in $\mathscr{F} = \mathscr{H}_K$, we have

$$\big|\phi(Yf(\mathbf{X})) - \phi(Yf'(\mathbf{X}))\big| \leq \big|f(\mathbf{X}) - f'(\mathbf{X})\big|,$$

Talagrand's contraction Lemma (Lemma 3.2 in Lecture 4) yields that,

$$\mathbb{E}\|\mathbf{Rad}_n(\phi \bullet f)\|_{\mathscr{H}_K} \leq \mathbb{E}\|\mathbf{Rad}_n(f)\|_{\mathscr{H}_K} \leq \lambda_n^{-1/2} K_0 \sqrt{\frac{1}{n}}.$$

Hyperparameter tuning

Then, by Corollary 3.1 in Lecture 6, if

$$\varepsilon_n \geq A_0 \lambda_n^{\frac{\alpha}{\alpha+1}} + \lambda_n^{-1/2} K_0 \sqrt{\frac{1}{n}} \geq \mathbf{Approx}_\phi(\lambda_n) + 8\mathbb{E}\|\mathbf{Rad}_n(\phi \bullet f)\|_{\mathscr{H}_K}.$$

Then,

$$\mathbb{P}\big(\mathscr{E}_\phi(\widehat{f}_n) \geq \varepsilon_n\big) \leq \exp\Big(-\frac{n\varepsilon_n^2}{8(1+5\varepsilon_n/6)}\Big).$$

We can tune λ_n to improve the convergence rate:

$$\varepsilon_n^* = \inf_{\lambda_n} A_0 \lambda_n^{\frac{\alpha}{1+\alpha}} + cK_0(n\lambda_n)^{-1/2} = O(n^{-\alpha/(1+3\alpha)}),$$

obtained by $\lambda_n = O(n^{-(\alpha+1)/(1+3\alpha)})$. Therefore, the convergence rate is given as:

$$\mathscr{E}(\widehat{f}_n) = O_P(\varepsilon_n^*) = O_P(n^{-\frac{\alpha}{1+3\alpha}}).$$

Remark 7.4 (How to improve?). (i) The trivial function class $\widehat{f}_n \in \{f \in \mathscr{H}_K : \|f\|_{\mathscr{H}_K} \leq \lambda_n^{-1/2}\}$ can be significantly improved by "low-noise" and "geometric noise" assumptions. (ii) Local/Random Rademacher complexity.

References

[Bartlett et al., 2006] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

[Cantelli, 1933] Cantelli, F. P. (1933). Sulla determinazione empirica delle leggi di probabilità. *Giornale dell'Istituto Italiano degli Attuari*, 4:421–424.

[Ciarlet, 2013] Ciarlet, P. G. (2013). *Linear and Nonlinear Functional Analysis with Applications*. SIAM.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

[Cox, 1972] Cox, D. R. (1972). The analysis of multivariate binary data. *Applied statistics*, pages 113–120.

[Fienberg, 2012] Fienberg, S. E. (2012). R.A. Fisher and the statistical abcs. *Statistical Science*, 27(3):300–306.

[Fisher, 1922] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A*, 222:309–368.

[Fisher, 1925] Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5):700–725.

[Freund and Schapire, 1997] Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

[Geisser, 1980] Geisser, S. (1980). Growth and decline of the basic component. *The American Statistician*, 34(3):129–135.

[Glivenko, 1933] Glivenko, V. (1933). Sulla determinazione empirica delle leggi di probabilità. *Giornale dell'Istituto Italiano degli Attuari*, 4:92–99.

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

[Kallianpur, 1955] Kallianpur, G. (1955). On Fisher's lower bound to asymptotic variance of a consistent estimate. *Sankhyā: The Indian Journal of Statistics*, 15:331–342.

[Lin, 2004] Lin, Y. (2004). A note on margin-based loss functions in classification. *Statistics & probability letters*, 68(1):73–82.

[Pedregosa, 2014] Pedregosa, F. (2014). Surrogate loss functions in machine learning.

[Rao, 1962] Rao, C. R. (1962). Apparent anomalies and irregularities in maximum likelihood estimation. *Sankhyā: The Indian Journal of Statistics, Series B*, 24:73–102.

[Savage, 1976] Savage, L. J. (1976). On rereading R.A. Fisher. *The Annals of Statistics*, 4(3):441–500.

[Steinwart and Scovel, 2007] Steinwart, I. and Scovel, C. (2007). Fast rates for support vector machines using gaussian kernels. *The Annals of Statistics*, 35(2):575–607.

[Zhang, 2004] Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85.