# Support Vector Machines

## Chap.3.1 - 3.5

Ingo Steinward, Andreas Christmann

Presenter: Sarah Kim

2018.02.19

# Contents

## Overview

- In many case, the loss describing a learning problem is not suitable when designing a learning algorithm.
- To reslove this issue, use a surrogate loss in the algorithm design.
- Goal of chaper 3: systematically develop a theory that makes it possible to identify suitable surrogate losses for general learning problem.

## Introduction

- Given a target loss, what surrogate loss is appropriate?
- Let $L_{\text{tar}}$ be a target loss that describes our learning goal and $L_{\text{sur}}$ be a surrogate loss. Given a loss function $L$ and a distribution $P$ on $X \times Y$, the $L$-risk of a measurable function $f\colon X \to \mathbb{R}$ is given by

$$\mathcal{R}_{L,P}(f) = \int_X \int_Y L(x, y, f(x))\, dP(y|x)\, dP_X(x).$$

# Introduction

- Question 3.1. Given a target loss $L_{\text{tar}}$, which surrogate losses $L_{\text{sur}}$ ensure the implication

$$\lim_{n\to\infty} \mathcal{R}_{L_{\text{sur}},P}(f_n) = \mathcal{R}^*_{L_{\text{sur}},P} \Rightarrow \lim_{n\to\infty} \mathcal{R}_{L_{\text{tar}},P}(f_n) = \mathcal{R}^*_{L_{\text{tar}},P} \tag{3.3}$$

for all sequences $(f_n)$ of measurable functions $f_n : X \to \mathbb{R}$?

- Question 3.2. Does there exist an increasing function $\Upsilon : [0,\infty) \to [0,\infty)$ that is continuous at 0 with $\Upsilon(0) = 0$ s.t., for all measurable $f : X \to \mathbb{R}$, we have

$$\mathcal{R}_{L_{\text{tar}},P}(f) - \mathcal{R}^*_{L_{\text{tar}},P} \le \Upsilon(\mathcal{R}_{L_{\text{sur}},P}(f) - \mathcal{R}^*_{L_{\text{sur}},P})?$$

## 3.1 Inner Risks and the Calibration Function

- Recall that the $L$-risk of a measurable function $f: X \to \mathbb{R}$ is given by $\mathcal{R}_{L,P}(f) = \int_X \int_Y L(x, y, f(x))\, dP(y|x)\, dP_X(x)$.

- Definition 3.3. Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss and $Q$ be a distribution on $Y$. We define the **inner $L$-risks** of $Q$ by

$$\mathcal{C}_{L,Q,x}(t) := \int_Y L(x, y, t)\, dQ(y), \quad x \in X, t \in \mathbb{R}.$$

Furthermore, the **minimal inner $L$-risks** are denoted by

$$\mathcal{C}^*_{L,Q,x} := \inf_{t \in \mathbb{R}} \mathcal{C}_{L,Q,x}(t), \quad x \in X.$$

## 3.1 Inner Risks and the Calibration Function

- By the definition 3.3., we obtain

$$\mathcal{R}_{L,P}(f) = \int_X \mathcal{C}_{L,P(\cdot|x),x}(f(x))dP_X(x). \tag{3.5}$$

- Lemma 3.4 shows that the Bayes risk $\mathcal{R}^*_{L,P}$ can be achieved by minimizing the inner risks $\mathcal{C}_{L,P(\cdot|x),x}$, $x \in X$.

- Lemma 3.4. Let $X$ be a complete measurble space, $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss, and $P$ be a distribtuion on $X \times Y$. Then $x \mapsto \mathcal{C}^*_{L,P(\cdot|x),x}$ is measurable and we have

$$\mathcal{R}^*_{L,P} = \int_X \mathcal{C}^*_{L,P(\cdot|x),x}dP_X(x). \tag{3.6}$$

## 3.1 Inner Risks and the Calibration Function

- Now assume that $\mathcal{R}^*_{L,P} < \infty$. Then the **excess risk** $\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}$ is defined and can be computed by

$$\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} = \int_X \mathcal{C}_{L,P(\cdot|x),x}(f(x)) - \mathcal{C}^*_{L,P(\cdot|x),x} dP_X(x)$$

- Split the analysis of excess risk into:
  1. the analysis of the **inner excess risk** $\mathcal{C}_{L,P(\cdot|x),x}(f(x)) - \mathcal{C}^*_{L,P(\cdot|x),x}$, $x \in X$;
  2. the investigation of the integration w.r.t. $P_X$.

## 3.1 Inner Risks and the Calibration Function

- We write

$$\mathcal{M}_{L,Q,x}(\epsilon) := \{t \in \mathbb{R} : \mathcal{C}_{L,Q,x}(t) < \mathcal{C}^*_{L,Q,x} + \epsilon\}, \quad \epsilon \in [0, \infty],$$

  for the sets containing the $\epsilon$-**approximate minimizers** of $\mathcal{C}_{L,Q,x}(\cdot)$.

  And the set of **exact minimizers** is denoted by

$$\mathcal{M}_{L,Q,x}(0^+) := \bigcap_{\epsilon>0} \mathcal{M}_{L,Q,x}(\epsilon).$$

- Example 3.8. For a distribution $Q$ on $Y := \{-1, 1\}$, we denote $\eta = Q(\{1\})$. Recall that the standard binary classification loss is defined by $L(y, t) := \mathbf{1}_{(-\infty,0]}(y\,\mathrm{sign}\,t), y \in Y, t \in \mathbb{R}$. For this loss, we have

$$\mathcal{C}_{L,\eta}(t) = \eta \mathbf{1}_{(-\infty,0)}(t) + (1-\eta)\mathbf{1}_{[0,\infty)}(t), \quad \eta \in [0,1], t \in \mathbb{R}$$

$$\mathcal{C}^*_{L,\eta}(t) = \min\{\eta, 1-\eta\}.$$

Hence,

$$\mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} = |2\eta - 1| \cdot \mathbf{1}_{(-\infty,0]}\Big((2\eta-1)\mathrm{sign}\,t\Big), \quad \eta \in [0,1] \tag{3.9}$$

And we find

$$\mathcal{M}_{L,\eta}(\epsilon) = \begin{cases} \mathbb{R} & \text{if } \epsilon > |2\eta - 1| \\ \{t \in \mathbb{R} : (2\eta-1)\mathrm{sign}\,t > 0\} & \text{if } 0 < \epsilon \leq |2\eta - 1|. \end{cases}$$

## 3.1 Inner Risks and the Calibration Function

- Lemma 3.10 (Properties of minimizers).

  Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss and $Q$ be a distribution on $Y$. For $x \in X$ and $t \in \mathbb{R}$, we have

  1. $\mathcal{M}_{L,Q,x}(0) = \emptyset$.
  2. $\mathcal{M}_{L,Q,x}(\epsilon) \neq \emptyset$ for some $\epsilon \in (0, \infty]$ iff $\mathcal{C}^*_{L,Q,x} < \infty$.
  3. $\mathcal{M}_{L,Q,x}(\epsilon_1) \subset \mathcal{M}_{L,Q,x}(\epsilon_2)$ for all $0 \leq \epsilon_1 \leq \epsilon_2 \leq \infty$.
  4. $t \in \mathcal{M}_{L,Q,x}(0^+)$ iff $\mathcal{C}_{L,Q,x}(t) = \mathcal{C}^*_{L,Q,x}$ and $\mathcal{C}^*_{L,Q,x} < \infty$.
  5. $t \in \mathcal{M}_{L,Q,x}(\infty)$ iff $\mathcal{C}_{L,Q,x}(t) < \infty$.

## 3.1 Inner Risks and the Calibration Function

- To show that we can use the set $\mathcal{M}_{L,P(\cdot|x),x}(\cdot)$ to construct $L$-risk minimizers, consider the following lemmas. Here, we let $X$ be a complete measurable space, $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss and $P$ be a distribution on $X \times Y$.
- Lemma 3.11 (Existence of approximate minimizers).
  For $\epsilon \in (0, \infty]$, the followings are equivalent:
  1. $\mathcal{C}^*_{L,P(\cdot|x),x} < \infty$ for $P_X$-almost all $x \in X$.
  2. There exists a measurable $f : X \to \mathbb{R}$ s.t. $f(x) \in \mathcal{M}_{L,P(\cdot,x),x}(\epsilon)$ for $P_X$-almost all $x \in X$.
- Lemma 3.12 (Existence of exact minimizers).
  Let $P$ be a distribution on $X \times Y$ satisfying $\mathcal{R}^*_{L,P} < \infty$. Then the followings are equivalent:
  1. $\mathcal{M}^*_{L,P(\cdot|x),x}(0^+) \neq \emptyset$ for $P_X$-almost all $x \in X$.
  2. There exists a measurable $f^* : X \to \mathbb{R}$ s.t. $\mathcal{R}_{L,P}(f^*) = \mathcal{R}^*_{L,P}$.

## 3.1 Inner Risks and the Calibration Function

- To show that we can use the set $\mathcal{M}_{L,P(\cdot|x),x}(\cdot)$ to construct $L$-risk minimizers, consider the following lemmas. Here, we let $X$ be a complete measurable space, $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss and $P$ be a distribution on $X \times Y$.
- Lemma 3.11 (Existence of approximate minimizers).
  For $\epsilon \in (0, \infty]$, the followings are equivalent:
  1. $\mathcal{C}^*_{L,P(\cdot|x),x} < \infty$ for $P_X$-almost all $x \in X$.
  2. There exists a measurable $f : X \to \mathbb{R}$ s.t. $f(x) \in \mathcal{M}_{L,P(\cdot,x),x}(\epsilon)$ for $P_X$-almost all $x \in X$.
- Lemma 3.12 (Existence of exact minimizers).
  Let $P$ be a distribution on $X \times Y$ satisfying $\mathcal{R}^*_{L,P} < \infty$. Then the followings are equivalent:
  1. $\mathcal{M}^*_{L,P(\cdot|x),x}(0^+) \neq \emptyset$ for $P_X$-almost all $x \in X$.
  2. There exists a measurable $f^* : X \to \mathbb{R}$ s.t. $\mathcal{R}_{L,P}(f^*) = \mathcal{R}^*_{L,P}$.

## 3.1 Inner Risks and the Calibration Function

- Assume for a moment that we have $L_{\mathrm{tar}}, L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0, \infty)$ s.t.

$$\emptyset \neq \mathcal{M}_{L_{\mathrm{sur}}, P(\cdot|x), x}(0^+) \subset \mathcal{M}_{L_{\mathrm{tar}}, P(\cdot|x), x}(0^+), \quad x \in X. \tag{3.13}$$

  Then Lemmas 3.4, 3.12 show that we have

$$\mathcal{R}_{L_{\mathrm{sur}}, P}(f) = \mathcal{R}^*_{L_{\mathrm{sur}}, P} \Rightarrow \mathcal{R}_{L_{\mathrm{tar}}, P}(f) = \mathcal{R}^*_{L_{\mathrm{tar}}, P}. \tag{3.14}$$

- Many learning procedures are able to find approximate minimizers, hence we need an approximate version of (3.14).

- Definition 3.13. Let $L_{\text{tar}}, L_{\text{sur}} : X \times Y \times \mathbb{R} \to [0, \infty)$ be loss functions, $Q$ be a distribution on $Y$, and $x \in X$. Then we define the **calibration function** $\delta_{\max}(\cdot, Q, x) : [0, \infty] \to [0, \infty]$ of $(L_{\text{tar}}, L_{\text{sur}})$ by

$$\delta_{\max}(\epsilon, Q, x) := \begin{cases} \inf_{t \in \mathbb{R} / \mathcal{M}_{L_{\text{tar}}, Q, x}(\epsilon)} \mathcal{C}_{L_{\text{sur}}, Q, x}(t) - \mathcal{C}^*_{L_{\text{sur}}, Q, x} & \text{if } \mathcal{C}^*_{L_{\text{sur}}, Q, x} < \infty \\ \infty & \text{if } \mathcal{C}^*_{L_{\text{sur}}, Q, x} = \infty \end{cases}$$

for all $\epsilon \in [0, \infty]$.

## 3.2 Asymptotic Theory of Surrogate Losses

- Theorem 3.17 (Asymptotic calibration of risks).

  Let $X$ be a complete measurable space, $L_{\mathrm{tar}}, L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0, \infty)$ be losses and $P$ be a distribution on $X \times Y$ s.t. $\mathcal{R}^*_{L_{\mathrm{tar}},P} < \infty$ and $\mathcal{R}^*_{L_{\mathrm{sur}},P} < \infty$. Then

  $$x \mapsto \delta_{\max}(\epsilon, P(\cdot|x), x)$$

  is measurable for all $\epsilon \in [0, \infty]$. In addition, consider

  *i*) For all $\epsilon \in (0, \infty]$, we have $P_X(\{x \in X : \delta_{\max}(\epsilon, P(\cdot|x), x) = 0\}) = 0$.

  *ii*) For all $\epsilon \in (0, \infty]$, $\exists \delta > 0$ s.t., for all measurable function $f : X \to \mathbb{R}$, we have

  $$\mathcal{R}_{L_{\mathrm{sur}},P}(f) < \mathcal{R}^*_{L_{\mathrm{sur}},P} + \delta \Rightarrow \mathcal{R}_{L_{\mathrm{tar}},P}(f) < \mathcal{R}^*_{L_{\mathrm{tar}},P} + \epsilon \qquad (3.18)$$

  Then we have *ii*) $\Rightarrow$ *i*). Furthermore, *i*) $\Rightarrow$ *ii*) holds if there exists a $P_X$-integrable function $b : X \to [0, \infty)$ s.t., for all $x \in X$, $t \in \mathbb{R}$, we have

  $$\mathcal{C}_{L_{\mathrm{tar}},P(\cdot|x),x}(t) \leq \mathcal{C}^*_{L_{\mathrm{tar}},P(\cdot|x),x} + b(x). \qquad (3.19)$$

- Theorem 3.17 shows that 'an almost surely strictly positive calibration function $\delta_{\max}$' is necessary for having an implication of the form

$$\mathcal{R}_{L_{\mathrm{sur}},P}(f_n) \to \mathcal{R}^*_{L_{\mathrm{sur}},P} \Rightarrow \mathcal{R}_{L_{\mathrm{tar}},P}(f_n) \to \mathcal{R}^*_{L_{\mathrm{tar}},P} \tag{3.21}$$

for all sequences $(f_n)$ of measurble functions. If (3.19) holds, 'an almost surely strictly positive calibration function $\delta_{\max}$' is sufficient for (3.21).

## 3.2 Asymptotic Theory of Surrogate Losses

- Definition 3.18. Let $L_{\text{tar}}, L_{\text{sur}} : X \times Y \times \mathbb{R} \to [0, \infty)$ be losses and $\mathcal{Q}$ be a set of distribtuions on $Y$. We say that $L_{\text{sur}}$ is **$L_{\text{tar}}$-calibrated** w.r.t. $\mathcal{Q}$ if, for all $\epsilon \in (0, \infty]$, $Q \in \mathcal{Q}$, and $x \in X$, we have

$$\delta_{\text{max}}(\epsilon, Q, x) > 0.$$

- Note that $L_{\text{sur}}$ is $L_{\text{tar}}$-calibrated w.r.t. $\mathcal{Q}$ iff for all $\epsilon > 0$, $Q \in \mathcal{Q}$, and $x \in X$, $\exists \delta > 0$ with

$$\mathcal{M}_{L_{\text{sur}}, Q, x}(\delta) \subset \mathcal{M}_{L_{\text{tar}}, Q, x}(\epsilon). \tag{3.22}$$

## 3.2 Asymptotic Theory of Surrogate Losses

- Corollary 3.19. Let $X$ be a complete measurable space, $L_{\text{tar}}, L_{\text{sur}} : X \times Y \times \mathbb{R} \to [0, \infty)$ be losses and $\mathcal{Q}$ be a set of distributions on $Y$. If $L_{\text{tar}}$ is bounded, then the followings are equivalent:
  i) $L_{\text{sur}}$ is $L_{\text{tar}}$-calibrated w.r.t. $\mathcal{Q}$.
  ii) For all $\epsilon > 0$ and all distributions $P$ of type $\mathcal{Q}$ with $\mathcal{R}^*_{L_{\text{sur}},P} < \infty$, $\exists \delta > 0$ s.t., for all measurable $f : X \to \mathbb{R}$, we have

$$\mathcal{R}_{L_{\text{sur}},P}(f) < \mathcal{R}^*_{L_{\text{sur}},P} + \delta \Rightarrow \mathcal{R}_{L_{\text{tar}},P}(f) < \mathcal{R}^*_{L_{\text{tar}},P} + \epsilon$$

  * We say that a distribution $P$ on $X \times Y$ is of **type** $\mathcal{Q}$ if $P(\cdot|x) \in \mathcal{Q}$ for $P_X$-almost all $x \in X$.

## 3.2 Asymptotic Theory of Surrogate Losses

- In Example 3.16, for the classification loss (target loss) and the least squares or the hinge loss (the surrogate losses), then the corresponding calibration functions are strictly positive.
  1. For $\epsilon > |2\eta - 1|$, $\delta_{\mathsf{max}, L_{\mathsf{class}}, L}(\epsilon, \eta) = \infty$.
  2. For $0 < \epsilon \leq |2\eta - 1|$,

  $$\delta_{\mathsf{max}, L_{\mathsf{class}}, L}(\epsilon, \eta) = \begin{cases} (2\eta - 1)^2 & \text{if } L = L_{\mathsf{LS}} \\ |2\eta - 1| & \text{if } L = L_{\mathsf{hinge}} \end{cases}$$

  Hence, by Cor. 3.19, both loss functions are reasonable surrogates in an asymptotic sense.

## 3.3 Inequalities between Excess Risks

- For $f$ with $\epsilon := \mathcal{R}_{L_{\text{tar}},P}(f) - \mathcal{R}^*_{L_{\text{tar}},P} > 0$, we want to find a $\delta(\epsilon) > 0$ s.t.,

$$\delta\Big(\mathcal{R}_{L_{\text{tar}},P}(f) - \mathcal{R}^*_{L_{\text{tar}},P}\Big) \leq \mathcal{R}_{L_{\text{sur}},P}(f) - \mathcal{R}^*_{L_{\text{sur}},P}. \tag{3.23}$$

- Definition 3.20. Let $I \subset \mathbb{R}$ be an interval and $g : I \to [0, \infty]$ be a function. Then the **Fenchel-Legendre bi-conjugate** $g^{**} : I \to [0, \infty]$ of $g$ is the largest convex function $h : I \to [0, \infty]$ satisfying $h \leq g$. Moreover, we write $g^{**}(\infty) := \lim_{t \to \infty} g^{**}(t)$ if $I = [0, \infty)$.

## 3.3 Inequalities between Excess Risks

- Definition 3.21. Let $\mathcal{Q}$ be a set of distribtuions on $Y$. Then the **uniform calibration function** w.r.t. $\mathcal{Q}$ is defined by

$$\delta_{\mathsf{max}}(\epsilon, \mathcal{Q}) := \inf_{Q \in \mathcal{Q}, x \in X} \delta_{\mathsf{max}}(\epsilon, Q, x) \quad \epsilon \in [0, \infty].$$

And we say that $L_{\mathsf{sur}}$ is **uniformly** $L_{\mathsf{tar}}$**-calibrated** w.r.t $\mathcal{Q}$ if $\delta_{\mathsf{max}}(\epsilon, \mathcal{Q}) > 0, \forall \epsilon > 0$.

## 3.3 Inequalities between Excess Risks

- **Theorem 3.22 (Uniform calibration inequalities).**

  Let $X$ be a complete measurable space, $L_{\mathrm{tar}}, L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0, \infty)$ be losses, and $\mathcal{Q}$ be a set of distributions on $Y$. And let $\delta : [0, \infty] \to [0, \infty]$ be an increasing function s.t.

$$\delta_{\max}(\epsilon, \mathcal{Q}) \geq \delta(\epsilon), \quad \epsilon \in [0, \infty]. \tag{3.25}$$

  Then for all distributions $P$ of type $\mathcal{Q}$ satisfying $\mathcal{R}^*_{L_{\mathrm{tar}},P} < \infty$ and $\mathcal{R}^*_{L_{\mathrm{sur}},P} < \infty$ and all measurable $f : X \to \mathbb{R}$, we have

$$\delta^{**}_{B_f}\left(\mathcal{R}_{L_{\mathrm{tar}},P}(f) - \mathcal{R}^*_{L_{\mathrm{tar}},P}\right) \leq \mathcal{R}_{L_{\mathrm{sur}},P}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},P}, \tag{3.26}$$

  where $\delta^{**}_{B_f} : [0, B_f] \to [0, \infty]$ is the biconjugate of $\delta_{|[0,B_f]}$, and $B_f$ is the supremum of the excess inner target risk w.r.t. $f$ i.e.,
  $B_f := \left\| x \mapsto (\mathcal{C}_{L_{\mathrm{tar}},P(\cdot|x),x}(f(x)) - \mathcal{C}^*_{L_{\mathrm{tar}},P(\cdot|x),x}) \right\|_\infty$.

## 3.3 Inequalities between Excess Risks

- Example 3.23. Let $L$ be the least squares loss or the hinge loss, $\mathcal{Q}_Y$ be the set of all distributions on $Y := \{-1, 1\}$, and $L_{\text{class}}$ be the binary classification loss. Using Example 3.16, we obtain

$$\delta_{\max, L_{\text{class}}, L}(\epsilon, \mathcal{Q}_Y) = \inf_{\eta \in [0,1]} \delta_{\max, L_{\text{class}}, L}(\epsilon, \eta) = \inf_{|2\eta - 1| \geq \epsilon} \delta_{\max, L_{\text{class}}, L}(\epsilon, \eta)$$

for all $\epsilon > 0$. For the least squares loss, $\delta_{\max, L_{\text{class}}, L}(\epsilon, \mathcal{Q}_Y) = \epsilon^2$, hence for all measurable $f : X \to \mathbb{R}$, we have

$$\mathcal{R}_{L_{\text{class}}, P}(f) - \mathcal{R}^*_{L_{\text{class}}, P} \leq \sqrt{\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}}.$$

For the hinge loss, $\delta_{\max, L_{\text{class}}, L}(\epsilon, \mathcal{Q}_Y) = \epsilon, \forall \epsilon > 0$, we have Zhang's inequality.

## 3.3 Inequalities between Excess Risks

- Theorem 3.25 (General calibration inequalities).

  Let $X$ be a complete measurable space, $L_{\mathrm{tar}}, L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0, \infty)$ be losses, and $P$ be a distribution on $X \times Y$ s.t. $\mathcal{R}^*_{L_{\mathrm{tar}},P} < \infty$. Assume that there exist $p > 0$ and functions $b : X \to [0, \infty]$ and $\delta : [0, \infty) \to [0, \infty)$ s.t.

  $$\delta_{\max}(\epsilon, P(\cdot|x), x) \geq b(x)\delta(\epsilon), \quad \epsilon \geq 0, x \in X, \tag{3.28}$$

  and $b^{-1} \in L_p(P_X)$. Then, for $\bar{\delta} := \delta^{\frac{p}{p+1}} : [0, \infty) \to [0, \infty)$ and all measurable $f : X \to \mathbb{R}$, we have

  $$\bar{\delta}^{**}_{B_f}\left(\mathcal{R}_{L_{\mathrm{tar}},P}(f) - \mathcal{R}^*_{L_{\mathrm{tar}},P}\right) \leq \|b^{-1}\|^{\frac{p}{p+1}}_{L_p(P_X)} \left(\mathcal{R}_{L_{\mathrm{sur}},P}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},P}\right)^{\frac{p}{p+1}}$$

  where $\bar{\delta}^{**}_{B_f} : [0, B_f] \to [0, \infty]$ is the biconjugate of $\bar{\delta}_{|[0,B_f]}$.

## 3.3 Inequalities between Excess Risks

- Our last goal in this section is to improve the previous inequalities for following type of loss.
- Definition 3.26. Let $A \subset X \times \mathbb{R}$ and $h : X \to [0, \infty)$ be measurable. Then we call $L : X \times \mathbb{R} \to [0, \infty)$ a **detection loss** w.r.t. $(A, h)$ if

$$L(x, t) = \mathbf{1}_A(x, t)h(x), \quad x \in X, t \in \mathbb{R}.$$

- Every detection loss function is measurable and an unsupervised loss function.
- For $x \in X$ and $t \in \mathbb{R}$, we have

$$\mathcal{C}_{L,x}(t) - \mathcal{C}^*_{L,x} = \begin{cases} 0 & \text{if } A(x) := \{t' \in \mathbb{R} : (x, t') \in A\} = \mathbb{R} \\ \mathbf{1}_A(x, t)h(x) & \text{otherwise.} \end{cases} \tag{3.29}$$

# 3.3 Inequalities between Excess Risks

- Theorem 3.27 (Asymptotic calibration for detection losses).
  Let $X$ be a complete m'ble space and $L_{\mathrm{tar}} : X \times \mathbb{R} \to [0, \infty)$ be a detection loss w.r.t. some $(A, h)$. And let $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss and $\mathcal{Q}$ be a set of distributions on $Y$. Then the followings are equivalent:

  *i*) $L_{\mathrm{sur}}$ is $L_{\mathrm{tar}}$-calibrated w.r.t. $\mathcal{Q}$.

  *ii*) For all $P$ of type $\mathcal{Q}$ that satisfy $h \in L_1(P_X)$ and $\mathcal{R}^*_{L_{\mathrm{sur}},P} < \infty$ and all $\epsilon > 0$, $\exists \delta > 0$ s.t. for all m'ble $f : X \to \mathbb{R}$ we have

$$\mathcal{R}_{L_{\mathrm{sur}},P}(f) < \mathcal{R}^*_{L_{\mathrm{sur}},P} + \delta \Rightarrow \mathcal{R}_{L_{\mathrm{tar}},P}(f) < \mathcal{R}^*_{L_{\mathrm{tar}},P} + \epsilon$$

## 3.3 Inequalities between Excess Risks

- **Theorem 3.28 (Calibration inequalities for detection losses).**
  Let $X$ be a complete m'ble space and $L_{\mathrm{tar}} : X \times \mathbb{R} \to [0, \infty)$ be a detection loss w.r.t. $(A, h)$, $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss and $P$ be a distributions on $X \times Y$ with $\mathcal{R}^*_{L_{\mathrm{tar}},P} < \infty$ and $\mathcal{R}^*_{L_{\mathrm{sur}},P} < \infty$. For $s > 0$, we write

$$B(s) := \left\{ x \in X : A(x) \neq \mathbb{R} \text{ and } \delta_{\max}(h(x), P(\cdot|x), x) < sh(x) \right\}.$$

If there exist constants $c > 0$ and $\alpha > 0$ s.t.

$$\int_X \mathbf{1}_{B(s)} h dP_X \leq (cs)^{\alpha}, \quad s > 0, \tag{3.30}$$

then for all m'ble functions $f : X \to \mathbb{R}$, we have

$$\mathcal{R}_{L_{\mathrm{tar}},P}(f) - \mathcal{R}^*_{L_{\mathrm{tar}},P} \leq 2c^{\frac{\alpha}{\alpha+1}} \left( \mathcal{R}_{L_{\mathrm{sur}},P}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},P} \right)^{\frac{\alpha}{\alpha+1}}.$$

- Remark 3.29.

For detection losses with $h = \mathbf{1}_X$, Thm. 3.28 yields an improvement over Thm. 3.25. For $\delta(\epsilon) = \epsilon^q$ and a $b : X \to [0, \infty]$ with $b^{-1} \in L_p(P_X)$ and $q \geq \frac{p+1}{p}$, then Thm. 3.25 gives

$$\mathcal{R}_{L_{\mathrm{tar}},P}(f) - \mathcal{R}^*_{L_{\mathrm{tar}},P} \leq \|b^{-1}\|^{1/q}_{L_p(P_X)} (\mathcal{R}_{L_{\mathrm{sur}},P}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},P})^{1/q}. \tag{3.31}$$

On the other hand, Thm. 3.28 yields

$$\mathcal{R}_{L_{\mathrm{tar}},P}(f) - \mathcal{R}^*_{L_{\mathrm{tar}},P} \leq 2\|b^{-1}\|^{\frac{p}{p+1}}_{L_p(P_X)} (\mathcal{R}_{L_{\mathrm{sur}},P}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},P})^{\frac{p}{p+1}}. \tag{3.32}$$

Since $q \geq \frac{p+1}{p}$, (3.32) is shaper than (3.31) if $\mathcal{R}_{L_{\mathrm{sur}},P}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},P}$ is sufficiently small.

## 3.4 Surrogates for Unweighted Binary Classification

- In this sub-section, $Y := \{-1, 1\}$, and we write $\mathcal{Q}_Y$ for the set of all distributions on $Y$, $\eta = Q(\{1\})$, for $Q \in \mathcal{Q}_Y$. If $L : Y \times \mathbb{R} \to [0, \infty)$ is a supervised loss, we use the notations $\mathcal{C}_{L,\eta}(t) := \mathcal{C}_{L,Q}(t)$, $t \in \mathbb{R}$, $\mathcal{C}^*_{L,\eta} := \mathcal{C}^*_{L,Q}$, as well as $\mathcal{M}_{L,\eta}(0^+) := \mathcal{M}_{L,Q}(0^+)$, $\mathcal{M}_{L,\eta}(\epsilon) := \mathcal{M}_{L,Q}(\epsilon)$, and $\delta_{\max}(\epsilon, \eta) := \delta_{\max}(\epsilon, Q)$ for $\epsilon \in [0, \infty]$.
- For margin-based losses, we have the following symmetries:

$$\mathcal{C}_{L,\eta}(t) = \mathcal{C}_{L,1-\eta}(-t) \quad \text{and} \quad \mathcal{C}^*_{L,\eta} = \mathcal{C}^*_{L,1-\eta},$$

$$\mathcal{M}_{L,\eta}(\epsilon) = -\mathcal{M}_{L,1-\eta}(\epsilon) \quad \text{and} \quad \mathcal{M}_{L,\eta}(0^+) = -\mathcal{M}_{L,1-\eta}(0^+),$$

## 3.4 Surrogates for Unweighted Binary Classification

- Definition 3.31. A supervised loss function $L : Y \times \mathbb{R} \to [0, \infty)$ is said to be **(uniformly) classification calibrated** if it is (uniformly) $L_{\text{class}}$-calibrated w.r.t. $\mathcal{Q}_Y$.
- Lemma 3.33 (Alternative to the calibration function).
  Let $L : Y \times \mathbb{R} \to [0, \infty)$ be a margin-based loss and $H : [0, 1] \to [0, \infty)$ be defined by

$$H(\eta) := \inf_{t \in \mathbb{R}:(2\eta-1)t \leq 0} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta}, \quad \eta \in [0, 1]. \qquad (3.37)$$

  Then the followings are true:

  *i*) L is classification calibrated iff $H(\eta) > 0$ for all $\eta \neq 1/2$.
  *ii*) If L is continuous, we have $\delta_{\text{max}}(\epsilon, \eta) = H(\eta)$ for all $0 < \epsilon \leq |2\eta - 1|$.
  *iii*) H is continuous and satisfies $H(\eta) = H(1 - \eta)$, $\eta \in [0, 1]$, and $H(1/2) = 0$.

# 3.4 Surrogates for Unweighted Binary Classification

- **Theorem 3.34 (Classification calibration).**

  Let $L : Y \times \mathbb{R} \to [0, \infty)$ be a margin-based loss. Then the followings are equivalent:

  *i)* $L$ is classification calibrated.

  *ii)* $L$ is uniformly classification calibrated.

  Furthermore, for $H$ defined by (3.37) and $\delta : [0,1] \to [0, \infty)$ defined by

  $$\delta(\epsilon) := H\Big(\frac{1+\epsilon}{2}\Big), \quad [0,1],$$

  the bi-conjugates of $\delta$ and $\delta_{\mathrm{max}}(\cdot, \mathcal{Q}_Y)$ satisfy

  $$\delta^{**}(\epsilon) \leq \delta^{**}_{\mathrm{max}, L_{\mathrm{class}}, L}(\epsilon, \mathcal{Q}_Y), \quad \epsilon \in [0,1], \tag{3.39}$$

  and both quantities are actually equal if $L$ is continuous.

  Finally, if $L$ is classification calibrated, we have $\delta^{**}(\epsilon) > 0$ for all $\epsilon \in (0,1]$.

- Theorem 3.36 (Test for classification calibration).
  Let $L$ be a convex, margin-based loss represented by $\phi : \mathbb{R} \to [0, \infty)$. Then the followings are equivalent:
  *i*) $L$ is classification calibrated.
  *ii*) $\phi$ is differentiable at $0$ and $\phi'(0) < 0$.

  Furthermore, if $L$ is classification calibrated, then

$$\delta^{**}_{\mathrm{max}}(\epsilon, \mathcal{Q}_Y) = \phi(0) - \mathcal{C}^*_{L, \frac{\epsilon}{2}}, \quad \epsilon \in [0, 1]. \tag{3.41}$$

## 3.5 Surrogates for Weighted Binary Classification

- $$L_{\alpha\text{-class}}(y,t) = \begin{cases} 1-\alpha & \text{if } y=1 \text{ and } t<0 \\ \alpha & \text{if } y=-1 \text{ and } t\geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

  where $\alpha \in (0,1)$ is a fixed weighting parameter and $Y := \{-1,1\}$.

- Let $L$ be a margin-based loss represented by some $\phi : \mathbb{R} \to [0,\infty)$. For $\alpha \in (0,1)$, define the **$\alpha$-weighted version** $L_\alpha$ of $L$ by

  $$L_\alpha(y,t) := \begin{cases} (1-\alpha)\phi(t) & \text{if } y=1 \\ \alpha\phi(-t) & \text{if } y=-1, \end{cases} \qquad t \in \mathbb{R}.$$

- To this end, we wil use

$$w_\alpha(\eta) := (1-\alpha)\eta + \alpha(1-\eta)$$
$$\theta_\alpha(\eta) := \frac{(1-\alpha)\eta}{(1-\alpha)\eta + \alpha(1-\eta)},$$

for $\eta \in [0,1]$.

## 3.5 Surrogates for Weighted Binary Classification

- Theorem 3.39 (Weighted classification calibration).

  Let $L$ be a margin-based loss function and $\alpha \in (0,1)$. We define $H_\alpha : [0,1] \to [0,\infty)$ by

$$H_\alpha(\eta) := \inf_{t\in\mathbb{R}:(\eta-\alpha)t\le 0} \mathcal{C}_{L_\alpha,\eta}(t) - \mathcal{C}^*_{L_\alpha,\eta}, \quad \eta \in [0,1]. \tag{3.47}$$

  Then the followings are equivalent:

  *i)* $L_\alpha$ is uniformly $L_{\alpha\text{-class}}$-calibrated w.r.t. $\mathcal{Q}_Y$.
  *ii)* $L_\alpha$ is $L_{\alpha\text{-class}}$-calibrated w.r.t. $\mathcal{Q}_Y$.
  *iii)* $L$ is classification calibrated.
  *iv)* $H_\alpha(\eta) > 0$ for all $\eta \in [0,1]$ with $\eta \neq \alpha$.

  Furthermore, if $H$ is defined by (3.37) then for all $\eta \in [0,1]$, we have

$$H_\alpha(\eta) = w_\alpha(\eta) H(\theta_\alpha(\eta)). \tag{3.48}$$

## 3.5 Surrogates for Weighted Binary Classification

- Theorem 3.40 (Weighted uniform classification calibration).
  Let $L$ be a margin-based loss function and $\alpha \in (0,1)$. For $\alpha_{\text{max}} := \max\{\alpha, 1-\alpha\}$, we define

  $$\delta_\alpha(\epsilon) := \inf_{\substack{\eta \in [0,1] \\ |\eta - \alpha| \geq \epsilon}} H_\alpha(\eta), \quad \epsilon \in [0, \alpha_{\text{max}}],$$

  where $H_\alpha(\cdot)$ is defined by (3.47). Then for all $\epsilon \in [0, \alpha_{\text{max}}]$, we have

  $$\delta_\alpha^{**}(\epsilon) \leq \delta_{\text{max}, L_{\alpha\text{-class}}}^{**}(\epsilon, \mathcal{Q}_Y),$$

  and if $L$ is continuous, both quantities are actually equal.
- To compute $\delta_\alpha(\epsilon)$, we can use $H_\alpha(\eta) = w_\alpha(\eta) H(\theta_\alpha(\eta))$. Then $\delta_\alpha$ is a continuous function and is strictly positive on $(0, \alpha_{\text{max}}]$ if $L$ is classificaion calibrated.

## 3.5 Surrogates for Weighted Binary Classification

| Loss function | $H(\eta)$ | $H_\alpha(\eta)$ | $\delta_\alpha^{**}(\varepsilon)$ |
|---|---|---|---|
| Least squares | $(2\eta-1)^2$ | $\frac{(\eta-\alpha)^2}{\alpha+\eta-2\alpha\eta}$ | $\frac{\varepsilon^2}{2\alpha(1-\alpha)+\varepsilon(1-2\alpha)}$ |
| Hinge loss | $\lvert 2\eta-1 \rvert$ | $\lvert \eta-\alpha \rvert$ | $\varepsilon$ |
| Squared hinge | $(2\eta-1)^2$ | $\frac{(\eta-\alpha)^2}{\alpha+\eta-2\alpha\eta}$ | $\frac{\varepsilon^2}{2\alpha(1-\alpha)+\varepsilon(1-2\alpha)}$ |

Figure 1 : The functions $H$, $H_\alpha$, and $\delta_\alpha^{**}$ for some common margin-based losses. The values for $\delta_\alpha^{**}$ are only for $\alpha$ with $0 < \alpha \leq 1/2$.

## 3.5 Surrogates for Weighted Binary Classification

- Theorem 3.41 (Using the correct weights).

  Let $\alpha, \beta \in (0, 1)$, $L$ be a margin-based, classification calibrated loss, and $L_\beta$ be its $\beta$-weighted version. Then $L_\beta$ is $L_{\alpha\text{-class}}$-calibrated iff $\beta = \alpha$.
- Using a weighted margin-based loss for unweighted classification problem may lead to methodical errors.