3

Surrogate Loss Functions (*)

Overview. *In many cases, the loss describing a learning problem is not suitable when designing a learning algorithm. A common approach to resolve this issue is to use a surrogate loss in the algorithm design. For example, we saw in the introduction that SVMs use the convex hinge loss instead of the discontinuous classification loss. The goal of this chapter is to systematically develop a theory that makes it possible to identify suitable surrogate losses for general learning problems.*

Prerequisites. *Besides Chapter 2 and Section A.3.3, only basic mathematical knowledge is required.*

Usage. *Sections 3.1 – 3.3 and 3.6 provide the theoretical framework required for Sections 3.4, 3.5, and 3.7 – 3.9, which deal with surrogate losses for common learning scenarios. These examples are important but not essential for classification, regression, and robustness, discussed in Chapters 8, 9, and 10, respectively. On the other hand, most of the material in this chapter is of general interest for machine learning and hence relatively independent of the rest of this book.*

In Chapter 2, we introduced some important learning scenarios and their corresponding loss functions. One way to design learning algorithms for these learning scenarios is to use a straightforward empirical risk minimization (ERM) ansatz based on the corresponding loss function. However, this approach may often be flawed, as the following examples illustrate:

- ERM optimization problems based on the classification loss are usually combinatorial problems, and even solving these problems approximately is often NP-hard.
- The least squares loss is known to be rather sensitive to outliers, and hence for certain data sets a (regularized) ERM approach based on this loss may fail, as we will see in Chapter 10.
- For some unsupervised learning scenarios, including the DLD scenario, we do not know the associated loss function since it depends on the unknown density.

These examples demonstrate that in many cases the loss function describing the learning problem is not suitable for a (regularized) ERM ansatz. Now recall that in the SVM approach discussed in the introduction one of the main ideas was to use the hinge loss function as a *surrogate* for the classification loss, and

consequently it is tempting to try surrogate losses in other learning scenarios, too. However, it is not hard to imagine that, given a *target loss*, not every loss function is a good surrogate, and hence we need some guidance for choosing a suitable surrogate loss.

Therefore, let us now describe what properties we do expect from good surrogate losses. To this end let, L_{tar} be a **target loss** that describes our learning goal and L_{sur} be a **surrogate loss**. Furthermore, assume that we have a learning method $\mathcal{A}$, e.g., a regularized L_{sur}-ERM approach, that asymptotically learns the surrogate learning problem defined by L_{sur}, i.e.,

$$\lim_{|D|\to\infty} \mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f_D) = \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} \tag{3.1}$$

holds in probability, where f_D is the decision function the method $\mathcal{A}$ produces for the training set D of length $|D|$. However, since our learning goal is defined by L_{tar}, we are actually interested in L_{tar}-consistency of $\mathcal{A}$, i.e., in

$$\lim_{|D|\to\infty} \mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f_D) = \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}}\,. \tag{3.2}$$

Obviously, we obtain the latter if the convergence in (3.1) implies the convergence in (3.2). This leads to the first question we will address in this chapter.

Question 3.1. *Given a target loss L_{tar}, which surrogate losses L_{sur} ensure the implication*

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

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

Question 3.1 is of purely asymptotic nature, i.e., it does consider any convergence rate in (3.1) or (3.2). Consequently, the surrogate losses that we find by answering Question 3.1 are a reasonable choice when dealing with consistency but may be less suitable when we wish to establish convergence rates for (3.2). This leads to the second question we will address.

Question 3.2. *Given a target loss L_{tar}, which surrogate losses L_{sur} allow us to deduce convergence rates for the right-hand side of (3.3) from convergence rates on the left-hand side of (3.3)?*

In particular, does there exist an increasing function $\Upsilon : [0,\infty) \to [0,\infty)$ that is continuous at 0 with $\Upsilon(0) = 0$ such that, for all measurable $f : X \to \mathbb{R}$, we have

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

Recall that we have already seen an example of such an inequality in Section 2.3, namely Zhang's inequality, which relates the excess classification risk to the excess hinge risk. In this chapter, we will systematically generalize the ideas used in the proof of that inequality to develop a general theory on

surrogate losses. The main results in this direction, including answers to the questions above, can be found in Sections 3.2 and 3.3. Furthermore, these general results will be applied to standard learning scenarios such as classification, regression, and density level detection in Sections 3.4, 3.5, 3.7, and 3.8.

3.1 Inner Risks and the Calibration Function

In order to address Questions 3.1 and 3.2, we need some tools and notions that will be introduced in this section. To this end let, us first recall that, given a loss function L and a distribution P on $X \times Y$, the L-risk of a measurable function $f : X \to \mathbb{R}$ is given by

$$\mathcal{R}_{L,\mathrm{P}}(f) = \int_X \int_Y L\big(x, y, f(x)\big)\, d\mathrm{P}(y|x)\, d\mathrm{P}_X(x). \tag{3.4}$$

Now, motivated by the calculations made in the proof of Zhang's inequality, the basic idea of our approach is to treat the inner and outer integrals *separately*. Besides some technical advantages, it will turn out that this approach has the important benefit of making our analysis rather independent of the specific distribution P, which, from the machine learning point of view, is unknown to us.

Let us begin with some fundamental definitions that will be used throughout this chapter.

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*

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

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

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

Finally, if L is a supervised loss, we usually drop the subscript x in these notations, and for unsupervised losses we analogously omit the subscript Q*.*

Note that by (3.4) and the definition of the inner risks, we immediately obtain

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

Our first goal is to establish the same relation between the minimal inner risks $\mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x}$, $x \in X$, and the Bayes risk $\mathcal{R}^*_{L,\mathrm{P}}$. To this end, we have to recall the notion of a complete measurable space given after Lemma A.3.3.

Lemma 3.4 (Computation of Bayes risks). *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 $X \times Y$. Then $x \mapsto \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x}$ is measurable and we have*

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

Proof. Let us define $\varphi : X \times \mathbb{R} \to [0, \infty]$ by

$$\varphi(x, t) := \mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(t) \,, \qquad\qquad x \in X,\, t \in \mathbb{R}.$$

Then φ is measurable by the measurability statement in Tonelli's theorem, and hence the first assertion follows from *iii)* of Lemma A.3.18 using $F(x) := \mathbb{R}$, $x \in X$. Consequently, the integral on the right-hand side of (3.6) exists, and it is easy to see that it satisfies

$$\mathcal{R}^*_{L,\mathrm{P}} = \inf_{f \in \mathcal{L}_0(X)} \int_X \mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(f(x)) \, d\mathrm{P}_X(x) \geq \int_X \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x} \, d\mathrm{P}_X(x) \,.$$

On the other hand, given $n \geq 1$, the second part of *iii)* in Lemma A.3.18 yields a measurable function $f_n : X \to \mathbb{R}$ with

$$\mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(f_n(x)) \;\leq\; \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x} + \frac{1}{n} \,, \qquad\qquad x \in X, \tag{3.7}$$

and hence we obtain

$$\mathcal{R}^*_{L,\mathrm{P}} \;\leq\; \mathcal{R}_{L,\mathrm{P}}(f_n) \;\leq\; \int_X \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x} \, d\mathrm{P}_X(x) + \frac{1}{n} \,.$$

Letting $n \to \infty$ then yields the assertion. □

Lemma 3.4 shows that the Bayes risk $\mathcal{R}^*_{L,\mathrm{P}}$ can be achieved by minimizing the inner risks $\mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(\,\cdot\,)$, $x \in X$, which in general will be easier than a direct minimization of $\mathcal{R}_{L,\mathrm{P}}(\,\cdot\,)$. Now assume that $\mathcal{R}^*_{L,\mathrm{P}} < \infty$. Then the **excess risk** $\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}$ is defined and can be computed by

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

for all measurable $f : X \to \mathbb{R}$. Consequently, we can split the analysis of the excess risk into:

i) the analysis of the **inner excess risks** $\mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(\,\cdot\,) - \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x}$, $x \in X$;
ii) the investigation of the integration with respect to P_X.

The benefit of this approach is that the analysis in *i)* only depends on P via the conditional distributions $\mathrm{P}(\,\cdot\,|x)$, and hence we can consider the excess inner risks $\mathcal{C}_{L,\mathrm{Q},x}(\,\cdot\,) - \mathcal{C}^*_{L,\mathrm{Q},x}$ for suitable classes of distributions Q on Y as a *template* for $\mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(\,\cdot\,) - \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x}$. This leads to the following definition.

Definition 3.5. *Let $\mathcal{Q}$ be a set of distributions on Y. We say that a distribution P on $X \times Y$ is of* ***type*** *$\mathcal{Q}$ if $\mathrm{P}(\,\cdot\,|x) \in \mathcal{Q}$ for P_X-almost all $x \in X$.*

In view of Questions 3.1 and 3.2, we are mainly interested in functions $f : X \to \mathbb{R}$ that almost minimize the risk under consideration. Following the idea of splitting the analysis into the steps *i)* and *ii)*, we therefore write

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

for the sets containing the **ε-approximate minimizers** of $\mathcal{C}_{L,\mathrm{Q},x}(\,\cdot\,)$. Moreover, the set of **exact minimizers** is denoted by

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

Again, for supervised and unsupervised losses, we usually omit the subscripts x and Q in the preceding definitions, respectively.

Before we investigate properties of the concepts above let us first illustrate these definitions with some examples. We begin with some margin-based losses introduced in Section 2.3. To this end, observe that any distribution Q on $Y := \{-1, 1\}$ can be uniquely described by an $\eta \in [0, 1]$ using the identification $\eta = \mathrm{Q}(\{1\})$. For a supervised loss $L : Y \times \mathbb{R} \to [0, \infty)$, we thus use the notation

$$\begin{aligned} \mathcal{C}_{L,\eta}(t) &:= \mathcal{C}_{L,\mathrm{Q}}(t)\,, \qquad t \in \mathbb{R}\,, \\ \mathcal{C}^*_{L,\eta} &:= \mathcal{C}^*_{L,\mathrm{Q}}\,, \end{aligned} \tag{3.8}$$

as well as $\mathcal{M}_{L,\eta}(0^+) := \mathcal{M}_{L,\mathrm{Q}}(0^+)$ and $\mathcal{M}_{L,\eta}(\varepsilon) := \mathcal{M}_{L,\mathrm{Q}}(\varepsilon)$ for $\varepsilon \in [0, \infty]$.

Example 3.6. Let L be the **least squares loss** defined in Example 2.26. For $t \in \mathbb{R}$ and $\eta \in [0, 1]$, a simple calculation then shows

$$\mathcal{C}_{L,\eta}(t) = \eta(1-t)^2 + (1-\eta)(1+t)^2 = 1 + 2t + t^2 - 4\eta t\,,$$

and hence elementary calculus gives $\mathcal{M}_{L,\eta}(0^+) = \{2\eta - 1\}$, $\mathcal{C}^*_{L,\eta} = 4\eta(1-\eta)$, and $\mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} = (t - 2\eta + 1)^2$ for all $t \in \mathbb{R}$ and $\eta \in [0, 1]$. ◁

Example 3.7. Recall that in Example 2.27 we defined the **hinge loss** by $L(y, t) := \max\{0, 1 - yt\}$, $y = \pm 1$, $t \in \mathbb{R}$. Now, for $\eta \in [0, 1]$ and $t \in \mathbb{R}$, a simple calculation shows

$$\mathcal{C}_{L,\eta}(t) = \begin{cases} \eta(1-t) & \text{if } t \le -1 \\ 1 + t(1-2\eta) & \text{if } t \in [-1, 1] \\ (1-\eta)(1+t) & \text{if } t \ge 1. \end{cases}$$

For $\eta \in [1/2, 1]$, we thus have

$$\mathcal{M}_{L,\eta}(0^+) = \begin{cases} [-1, 1] & \text{if } \eta = \frac{1}{2} \\ \{1\} & \text{if } \frac{1}{2} < \eta < 1 \\ [1, \infty) & \text{if } \eta = 1\,, \end{cases}$$

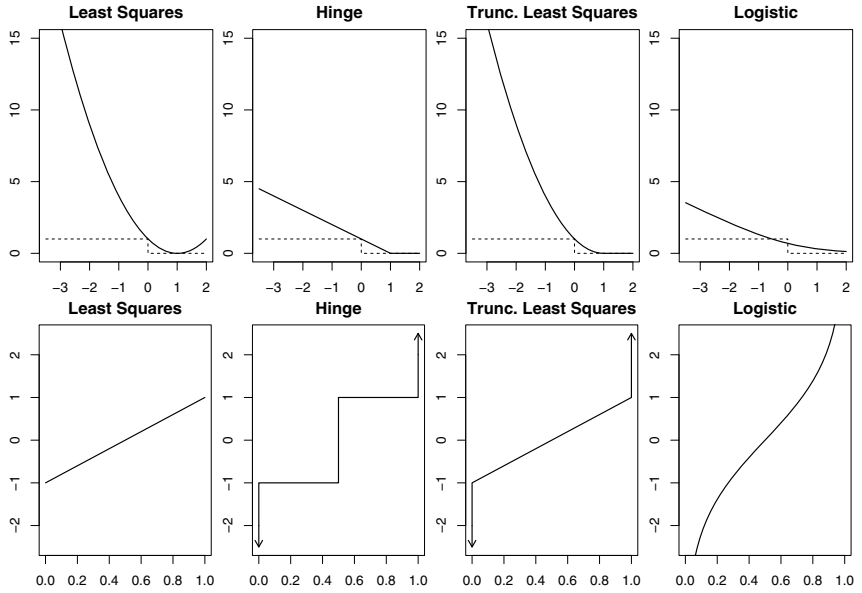

Fig. 3.1. The representing functions φ for some important margin-based loss functions L (top row) and their minimizing sets $\mathcal{M}_{L,\eta}(0^+)$, $\eta \in [0,1]$ (bottom row). For some losses and values of η, these sets are not singletons. This situation is displayed by vertical lines. Moreover, the arrows at the ends of some of these vertical lines indicate that the corresponding set is unbounded in the direction of the arrow.

$\mathcal{C}^*_{L,\eta} = 2(1-\eta)$, and

$$\mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} = \begin{cases} 3\eta - 2 - \eta t & \text{if } t \le -1 \\ (1-t)(2\eta - 1) & \text{if } t \in [-1,1] \\ (t-1)(1-\eta) & \text{if } t \ge 1. \end{cases}$$

In addition, similar formulas hold for $\eta \in [0, 1/2]$ by symmetry. ◁

Both margin-based loss functions discussed above will serve us as surrogates for the classification loss. Therefore, let us now consider the inner risks and the set of minimizers for the standard classification loss itself.

Example 3.8. Recall that the standard **binary classification loss** is defined by $L(y,t) := \mathbf{1}_{(-\infty,0]}\big(y \operatorname{sign} t\big)$, $y \in Y$, $t \in \mathbb{R}$. For this loss, the inner risk is given by

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

for all $\eta \in [0,1]$ and $t \in \mathbb{R}$. From this we easily conclude $\mathcal{C}^*_{L,\eta} = \min\{\eta, 1-\eta\}$, which in turn yields

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

for all $\eta \in [0,1]$ and $t \in \mathbb{R}$. Considering the cases $\varepsilon > |2\eta - 1|$ and $\varepsilon \leq |2\eta - 1|$ separately, we thus find

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

Let us finally determine the inner risks and their minimizers for a more elaborate example.

Proposition 3.9 (Quantiles and the pinball loss). *For $\tau \in (0,1)$, let L be the τ-pinball loss defined in Example 2.43. Moreover, let Q be a distribution on $\mathbb{R}$ with $|\mathrm{Q}|_1 < \infty$ and let t^* be a τ-quantile of Q, i.e., we have*

$$\mathrm{Q}\big((-\infty, t^*]\big) \geq \tau \qquad \text{and} \qquad \mathrm{Q}\big([t^*, \infty)\big) \geq 1 - \tau\,.$$

Then there exist real numbers $q_+, q_- \geq 0$ such that $q_+ + q_- = \mathrm{Q}(\{t^*\})$ and

$$\mathcal{C}_{L,\mathrm{Q}}(t^* + t) - \mathcal{C}^*_{L,\mathrm{Q}} = tq_+ + \int_0^t \mathrm{Q}\big((t^*, t^* + s)\big)\, ds\,, \tag{3.10}$$

$$\mathcal{C}_{L,\mathrm{Q}}(t^* - t) - \mathcal{C}^*_{L,\mathrm{Q}} = tq_- + \int_0^t \mathrm{Q}\big((t^* - s, t^*)\big)\, ds\,, \tag{3.11}$$

for all $t \geq 0$. Moreover, we have

$$\mathcal{M}_{L,\mathrm{Q}}(0^+) = \{t^*\} \cup \big\{t > t^* \colon q_+ + \mathrm{Q}((t^*, t)) = 0\big\} \cup \big\{t < t^* \colon q_- + \mathrm{Q}((-t, t^*)) = 0\big\}.$$

Proof. Recall from Example 2.43 that distance-based τ-pinball loss is represented by

$$\psi(r) = \begin{cases} (\tau - 1)r, & \text{if } r < 0 \\ \tau r, & \text{if } r \geq 0. \end{cases}$$

Now let us consider the distribution $\mathrm{Q}^{(t^*)}$ defined by $\mathrm{Q}^{(t^*)}(A) := \mathrm{Q}(t^* + A)$ for all measurable $A \subset \mathbb{R}$. Then it is not hard to see that 0 is a τ-quantile of $\mathrm{Q}^{(t^*)}$. Moreover, we obviously have $\mathcal{C}_{L,\mathrm{Q}}(t^* + t) = \mathcal{C}_{L,\mathrm{Q}^{(t^*)}}(t)$, and hence we may assume without loss of generality that $t^* = 0$. Then our assumptions together with $\mathrm{Q}((-\infty, 0]) + \mathrm{Q}([0, \infty)) = 1 + \mathrm{Q}(\{0\})$ yield $\tau \leq \mathrm{Q}((-\infty, 0]) \leq \tau + \mathrm{Q}(\{0\})$, i.e., there exists a q_+ satisfying $0 \leq q_+ \leq \mathrm{Q}(\{0\})$ and

$$\mathrm{Q}((-\infty, 0]) = \tau + q_+\,. \tag{3.12}$$

Let us now prove the first expression for the inner risks of L. To this end, we first observe that for $t \geq 0$ we have

$$\int_{y<t} (y - t)\, d\mathrm{Q}(y) = \int_{y<0} y\, d\mathrm{Q}(y) - t\mathrm{Q}((-\infty, t)) + \int_{0 \leq y < t} y\, d\mathrm{Q}(y)$$

and

$$\int_{y\geq t}(y-t)\,d\mathrm{Q}(y) = \int_{y\geq 0} y\,d\mathrm{Q}(y) - t\mathrm{Q}([t,\infty)) - \int_{0\leq y<t} y\,d\mathrm{Q}(y)\,.$$

Consequently, we obtain

$$\begin{aligned}\mathcal{C}_{L,\mathrm{Q}}(t) &= (\tau-1)\int_{y<t}(y-t)\,d\mathrm{Q}(y) + \tau\int_{y\geq t}(y-t)\,d\mathrm{Q}(y)\\ &= \mathcal{C}_{L,\mathrm{Q}}(0) - \tau t + t\mathrm{Q}((-\infty,0)) + t\mathrm{Q}([0,t)) - \int_{0\leq y<t} y\,d\mathrm{Q}(y)\,.\end{aligned}$$

Moreover, using Lemma A.3.11, we find

$$\begin{aligned}t\,\mathrm{Q}([0,t)) - \int_{0\leq y<t} y\,d\mathrm{Q}(y) &= \int_0^t \mathrm{Q}([0,t))\,ds - \int_0^t \mathrm{Q}([s,t))\,ds\\ &= t\,\mathrm{Q}(\{0\}) + \int_0^t \mathrm{Q}((0,s))\,ds\,,\end{aligned}$$

and since (3.12) implies $\mathrm{Q}((-\infty,0))+\mathrm{Q}(\{0\}) = \mathrm{Q}((-\infty,0]) = \tau+q_+$, we thus obtain

$$\mathcal{C}_{L,\mathrm{Q}}(t) = \mathcal{C}_{L,\mathrm{Q}}(0) + tq_+ + \int_0^t \mathrm{Q}\big((0,s)\big)\,ds\,.$$

Moreover, applying this equation to the pinball loss with parameter $1-\tau$ and the distribution $\bar{\mathrm{Q}}$ defined by $\bar{\mathrm{Q}}(A) := \mathrm{Q}(-A)$, $A\subset\mathbb{R}$ measurable, gives a real number $0\leq q_-\leq \mathrm{Q}(\{0\})$ such that $\mathrm{Q}([0,\infty)) = 1-\tau+q_-$ and

$$\mathcal{C}_{L,\mathrm{Q}}(-t) = \mathcal{C}_{L,\mathrm{Q}}(0) + tq_- + \int_0^t \mathrm{Q}\big((-s,0)\big)\,ds$$

for all $t\geq 0$. Consequently, $t^* = 0$ is a minimizer of $\mathcal{C}_{L,\mathrm{Q}}(\,\cdot\,)$ and hence we find both (3.10) and (3.11). Moreover, combining $\mathrm{Q}([0,\infty)) = 1-\tau+q_-$ with (3.12), we find $q_+ + q_- = \mathrm{Q}(\{0\})$. Finally, the formula for the set of exact minimizers is an obvious consequence of (3.10) and (3.11). □

Let us now return to our general theory. We begin with the following lemma, which collects some useful properties of the sets $\mathcal{M}_{L,\mathrm{Q},x}(\,\cdot\,)$. Its proof is left as an exercise.

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 then have:*

i) $\mathcal{M}_{L,\mathrm{Q},x}(0) = \emptyset$.
ii) $\mathcal{M}_{L,\mathrm{Q},x}(\varepsilon)\neq\emptyset$ *for some* $\varepsilon\in(0,\infty]$ *if and only if* $\mathcal{C}^*_{L,\mathrm{Q},x} < \infty$.
iii) $\mathcal{M}_{L,\mathrm{Q},x}(\varepsilon_1)\subset\mathcal{M}_{L,\mathrm{Q},x}(\varepsilon_2)$ *for all* $0\leq\varepsilon_1\leq\varepsilon_2\leq\infty$.
iv) $t\in\mathcal{M}_{L,\mathrm{Q},x}(0^+)$ *if and only if* $\mathcal{C}_{L,\mathrm{Q},x}(t) = \mathcal{C}^*_{L,\mathrm{Q},x}$ *and* $\mathcal{C}^*_{L,\mathrm{Q},x} < \infty$.
v) $t\in\mathcal{M}_{L,\mathrm{Q},x}(\infty)$ *if and only if* $\mathcal{C}_{L,\mathrm{Q},x}(t) < \infty$.

Our goal in the following two lemmas is to show that we can use the sets $\mathcal{M}_{L,\mathrm{P}(\,\cdot\,|x),x}(\,\cdot\,)$ to construct (approximate) L-risk minimizers. Note that the main difficulty in these lemmas is to ensure the measurability of the (approximate) minimizers.

Lemma 3.11 (Existence of approximate minimizers). *Let X be a complete measurable space, $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, P be a distribution on $X \times Y$, and $\varepsilon \in (0,\infty]$. Then the following statements are equivalent:*

i) $\mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x} < \infty$ *for* P_X*-almost all* $x \in X$.
ii) There exists a measurable $f : X \to \mathbb{R}$ *such that* $f(x) \in \mathcal{M}_{L,\mathrm{P}(\,\cdot\,|x),x}(\varepsilon)$ *for* P_X*-almost all* $x \in X$.

Proof. *ii)* $\Rightarrow$ *i)*. This immediately follows from *ii)* of Lemma 3.10.

i) $\Rightarrow$ *ii)*. Let $n \geq 1$ with $1/n < \varepsilon$. As in the proof of Lemma 3.4, we then obtain a measurable function $f_n : X \to \mathbb{R}$ satisfying (3.7) for all $x \in X$. Since $\mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x} < \infty$ for P_X-almost all $x \in X$, we thus find the assertion. □

While the preceding lemma characterizes the situations where *uniform* ε-approximate minimizers exist, the following lemma characterizes L-risks that have an *exact* minimizer, i.e., a Bayes decision function.

Lemma 3.12 (Existence of exact minimizers). *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 $X \times Y$ satisfying $\mathcal{R}^*_{L,\mathrm{P}} < \infty$. Then the following are equivalent:*

i) $\mathcal{M}_{L,\mathrm{P}(\,\cdot\,|x),x}(0^+) \neq \emptyset$ *for* P_X*-almost all* $x \in X$.
ii) There exists a measurable $f^* : X \to \mathbb{R}$ *such that* $\mathcal{R}_{L,\mathrm{P}}(f^*) = \mathcal{R}^*_{L,\mathrm{P}}$.

Moreover, if one of the conditions is satisfied, every Bayes decision function $f^*_{L,\mathrm{P}} : X \to \mathbb{R}$ *satisfies* $f^*_{L,\mathrm{P}}(x) \in \mathcal{M}_{L,\mathrm{P}(\,\cdot\,|x),x}(0^+)$ *for* P_X*-almost all* $x \in X$.

Proof. *i)* $\Rightarrow$ *ii)*. Let φ and F be defined as in the proof of Lemma 3.4. Using the last part of *iii)* in Lemma A.3.18, we then find a measurable $f^* : X \to \mathbb{R}$ with $f^*(x) \in \mathcal{M}_{L,\mathrm{P}(\,\cdot\,|x),x}(0^+)$ for P_X-almost all $x \in X$. Obviously, part *iv)* of Lemma 3.10 and Lemma 3.4 then show $\mathcal{R}_{L,\mathrm{P}}(f^*) = \mathcal{R}^*_{L,\mathrm{P}}$.

ii) $\Rightarrow$ *i)*. Let $f^*_{L,\mathrm{P}}$ be a Bayes decision function, i.e., it satisfies $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) = \mathcal{R}^*_{L,\mathrm{P}}$. Since $\mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(f^*_{L,\mathrm{P}}) \geq \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x}$ for all $x \in X$, Lemma 3.4 together with $\mathcal{R}^*_{L,\mathrm{P}} < \infty$ then yields $\mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(f^*_{L,\mathrm{P}}) = \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x}$ for P_X-almost all $x \in X$. Moreover, $\mathcal{R}^*_{L,\mathrm{P}} < \infty$ implies $\mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x} < \infty$ for P_X-almost all $x \in X$, and hence we find $f^*_{L,\mathrm{P}}(x) \in \mathcal{M}_{L,\mathrm{P}(\,\cdot\,|x),x}(0^+)$ for P_X-almost all $x \in X$ by part *iv)* of Lemma 3.10. □

Let us now assume for a moment that we have two loss functions $L_{\mathrm{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ such that

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

Then Lemmas 3.4 and 3.12 show that every exact minimizer of $\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(\,\cdot\,)$ is also an exact minimizer of $\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(\,\cdot\,)$, i.e., we have the implication

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

However, exact minimizers do not necessarily exist, as one can see by combining Lemma 3.11 with Lemma 3.12, and even if they do exist, it is rather unlikely that we will find them by a learning procedure. On the other hand, we have already indicated in Chapter 1 that many learning procedures are able to find approximate minimizers, and therefore we need an *approximate* version of (3.14) to answer Question 3.1. Now, the key idea for establishing such a modification of (3.14) is to consider approximate versions of (3.13). To this end, we begin with the following fundamental definition.

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

$$\delta_{\max}(\varepsilon,\mathrm{Q},x) := \begin{cases} \inf\limits_{\substack{t\in\mathbb{R}\\ t\notin\mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)}} \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} & \text{if } \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} < \infty \\ \infty & \text{if } \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} = \infty \end{cases}$$

for all $\varepsilon \in [0,\infty]$. Moreover, we write $\delta_{\max,L_{\mathrm{tar}},L_{\mathrm{sur}}}(\varepsilon,\mathrm{Q},x) := \delta_{\max}(\varepsilon,\mathrm{Q},x)$ whenever it is necessary to explicitly mention the target and surrogate losses. Finally, if both losses are supervised, we usually omit the argument x.

The following lemma collects some simple though extremely important properties of the calibration function.

Lemma 3.14 (Properties of the calibration function). *Let $L_{\mathrm{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be losses and Q be a distribution on Y. For all $x \in X$ and $\varepsilon \in [0,\infty]$, we then have:*

i) $\mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(\delta_{\max}(\varepsilon,\mathrm{Q},x)) \subset \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$.
ii) $\mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(\delta) \not\subset \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$ whenever $\delta > \delta_{\max}(\varepsilon,\mathrm{Q},x)$.

Consequently, the calibration function can be computed by

$$\delta_{\max}(\varepsilon,\mathrm{Q},x) = \max\bigl\{\delta \in [0,\infty] : \mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(\delta) \subset \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)\bigr\}\,. \tag{3.15}$$

Finally, if both $\mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{Q},x} < \infty$ and $\mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} < \infty$, then for all $t \in \mathbb{R}$ we have

$$\delta_{\max}\bigl(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{Q},x},\mathrm{Q},x\bigr) \;\le\; \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x}\,. \tag{3.16}$$

Inequality (3.16) will be the key ingredient when we compare the excess L_{tar}-risk with the excess L_{sur}-risk since it allows us to compare the inner integrals of these risks. Furthermore, one can show by *ii)* that the calibration function is the optimal way to compare these inner integrals. We refer to Exercise 3.3 for details.

Proof. Let us first assume $\mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} = \infty$. Then we have $\delta_{\max}(\varepsilon, \mathrm{Q}, x) = \infty$ and hence *ii)* is trivially satisfied. Moreover, we have $\mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(\delta_{\max}(\varepsilon, \mathrm{Q}, x)) = \emptyset$ by *ii)* of Lemma 3.10, and hence we obtain *i)*. Let us now assume $\mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} < \infty$. Then, for $t \in \mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(\delta_{\max}(\varepsilon, \mathrm{Q}, x))$, we have

$$\mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} < \delta_{\max}(\varepsilon, \mathrm{Q}, x) = \inf_{\substack{t' \in \mathbb{R} \\ t' \notin \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)}} \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t') - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x},$$

which shows $t \in \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$. For the proof of the second assertion, let us fix a δ with $\delta > \delta_{\max}(\varepsilon, \mathrm{Q}, x)$. By definition, this means

$$\inf_{\substack{t \in \mathbb{R} \\ t \notin \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)}} \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} = \delta_{\max}(\varepsilon, \mathrm{Q}, x) < \delta,$$

and hence there exists a $t \in \mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(\delta)$ with $t \notin \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$. This shows part *ii)*. Moreover, (3.15) is a direct consequence of *i)* and *ii)*.

Let us finally prove Inequality (3.16). To this end, we fix a $t \in \mathbb{R}$ and write $\varepsilon := \mathcal{C}_{L_{\mathrm{tar}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{Q},x}$. Then have $t \notin \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$, which implies $t \notin \mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(\delta_{\max}(\varepsilon, \mathrm{Q}, x))$ by *i)*. The latter means

$$\begin{aligned}\mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) &\geq \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} + \delta_{\max}(\varepsilon, \mathrm{Q}, x) \\ &= \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} + \delta_{\max}\left(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{Q},x}, \mathrm{Q}, x\right). \end{aligned}$$

□

Due to algorithmic reasons, we are often interested in *convex* surrogate losses. For such surrogates, the calibration function can be easily computed.

Lemma 3.15 (Calibration function for convex surrogates). *Let* Q *be a distribution on* Y*,* $L_{\mathrm{tar}} : X \times Y \times \mathbb{R} \to [0, \infty)$ *be a loss, and* $x \in X$*,* $\varepsilon > 0$ *such that* $\mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$ *is an interval. Moreover, let* $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0, \infty)$ *be a convex loss such that* $\mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) < \infty$ *for all* $t \in \mathbb{R}$*. If* $\mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(0^+) \subset \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(0^+)$*, then we have*

$$\delta_{\max}(\varepsilon, \mathrm{Q}, x) = \min\left\{\mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t_\varepsilon^-), \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t_\varepsilon^+)\right\} - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x}, \tag{3.17}$$

where we used the definitions $t_\varepsilon^- := \inf \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$*,* $t_\varepsilon^+ := \sup \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$*, and* $\mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(\pm\infty) := \infty$*.*

Proof. Obviously, $\mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(\,\cdot\,) : \mathbb{R} \to [0, \infty)$ is convex, and thus it is also continuous by Lemma A.6.2. Since $\mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$ is an interval, we then obtain

$$\delta_{\max}(\varepsilon, \mathrm{Q}, x) = \min\Big\{\inf_{t \leq t_\varepsilon^-} \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t), \inf_{t \geq t_\varepsilon^+} \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t)\Big\} - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x}.$$

Moreover, for $t < t_\varepsilon^-$, we have $t \notin \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$ and hence $t \notin \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(0^+)$. From this we conclude that $t \notin \mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(0^+)$, i.e., $\mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) > \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x}$. Consequently, the convexity of $t \mapsto \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t)$ shows that this map is strictly decreasing on $(-\infty, t_\varepsilon^-]$, and hence we obtain $\inf\{\mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) : t \leq t_\varepsilon^-\} = \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t_\varepsilon^-)$. For $t \geq t_\varepsilon^+$, we can argue analogously. □

Let us close this section with an example that illustrates how to compute the calibration function for specific loss functions.

Example 3.16. Let L be either the **least squares loss** L_{LS} or the **hinge loss** L_{hinge}. We write L_{class} for the binary classification loss and identify distributions Q on $\{-1,1\}$ by $\eta := \mathrm{Q}(\{1\})$. Then Lemma 3.15 together with Example 3.8 yields $\delta_{\max,L_{\mathrm{class}},L}(\varepsilon,\eta) = \infty$ if $\varepsilon > |2\eta-1|$. Moreover, for $0 < \varepsilon \leq |2\eta-1|$, we find

$$\delta_{\max,L_{\mathrm{class}},L}(\varepsilon,\eta) = \mathcal{C}_{L,\eta}(0) - \mathcal{C}^*_{L,\eta} = \begin{cases} (2\eta-1)^2 & \text{if } L = L_{\mathrm{LS}} \\ |2\eta-1| & \text{if } L = L_{\mathrm{hinge}} \end{cases}$$

by applying Examples 3.6 and 3.7, respectively. In particular note that in both cases we have $\delta_{\max,L_{\mathrm{class}},L}(\varepsilon,\eta) > 0$ for all $\eta \in [0,1]$ and all $\varepsilon > 0$. ◁

3.2 Asymptotic Theory of Surrogate Losses

In this section, we investigate the asymptotic relationship between excess risks in the sense of Question 3.1. The main result in this direction is the following theorem.

Theorem 3.17 (Asymptotic calibration of risks). *Let X be a complete measurable space, $L_{\mathrm{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be losses, and P be a distribution on $X \times Y$ such that $\mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} < \infty$ and $\mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} < \infty$. Then*

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

is measurable for all $\varepsilon \in [0,\infty]$. In addition, consider the following statements:

i) For all $\varepsilon \in (0,\infty]$, we have $\mathrm{P}_X\big(\{x \in X : \delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) = 0\}\big) = 0$.
ii) For all $\varepsilon \in (0,\infty]$, there exists a $\delta > 0$ such that, for all measurable functions $f : X \to \mathbb{R}$, we have

$$\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) < \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} + \delta \quad \Longrightarrow \quad \mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) < \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} + \varepsilon\,. \tag{3.18}$$

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

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

Proof. To show the measurability of $\delta_{\max}(\,\cdot\,, \mathrm{P}(\,\cdot\,|x), x)$, we may assume without loss of generality that we have $\mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x} < \infty$ and $\mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x} < \infty$ for all $x \in X$. We equip $[0,\infty]$ with the Borel σ-algebra and write $A := [\varepsilon,\infty]$. Furthermore, let $h : X \times \mathbb{R} \to [0,\infty]$ be defined by

$$h(x,t) := \mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\,, \qquad (x,t) \in X \times \mathbb{R}\,.$$

Then h is measurable and, for the set-valued function $F : X \to 2^{\mathbb{R}}$ defined by $F(x) := \{t \in \mathbb{R} : h(x,t) \in A\}$, $x \in X$, we have $\mathbb{R}\backslash\mathcal{M}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}(\varepsilon) = F(x)$ for all $x \in X$. Furthermore, $\varphi : X \times \mathbb{R} \to [0,\infty]$ defined by

$$\varphi(x,t) := \mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\,, \qquad (x,t) \in X \times \mathbb{R}\,,$$

is measurable. Now, for all $x \in X$, our construction yields

$$\delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) = \inf_{t \in F(x)} \varphi(x,t)\,,$$

and hence $x \mapsto \delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x)$ is measurable by Lemma A.3.18.

ii) $\Rightarrow$ *i)*. Assume that *i)* is not true. Then there is an $\varepsilon \in (0,\infty]$ such that

$$B := \bigl\{x \in X : \delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) = 0 \text{ and } \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x} < \infty\bigr\}$$

satisfies $\mathrm{P}_X(B) > 0$. Note that for $x \in B$ we have $\mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x} < \infty$ by the very definition of the calibration function. In addition, for $x \in B$, we have $\delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) = 0$ and hence there exists a $t \in \mathbb{R}\backslash\mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$. Using the notation of the first part of the proof, this t satisfies $h(x,t) \geq \varepsilon$ and hence we have $F(x) \neq \emptyset$. This shows $B \subset \operatorname{Dom} F$. By Lemma A.3.18, there thus exist measurable functions $f_n^{(1)} : X \to \mathbb{R}$ such that

$$\mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\bigl(f_n^{(1)}(x)\bigr) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x} \;\leq\; \frac{1}{n}$$

and

$$\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\bigl(f_n^{(1)}(x)\bigr) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x} \;\geq\; \varepsilon$$

for all $x \in B$ and $n \geq 1$. Furthermore, by Lemma 3.11, we find measurable functions $f_n^{(2)} : X \to \mathbb{R}$, $n \geq 1$, with

$$\mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\bigl(f_n^{(2)}(x)\bigr) \;<\; \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x} + \frac{1}{n}$$

for P_X-almost all $x \in X$. We define $f_n : X \to \mathbb{R}$ by

$$f_n(x) := \begin{cases} f_n^{(1)}(x) & \text{if } x \in B \\ f_n^{(2)}(x) & \text{otherwise.} \end{cases}$$

Then f_n is measurable and our construction yields both

$$\begin{aligned} \mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f_n) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} &\geq \int_B \Bigl(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\bigl(f_n(x)\bigr) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\Bigr) d\mathrm{P}_X(x) \\ &\geq \varepsilon\, \mathrm{P}_X(B) \end{aligned}$$

and $\lim_{n\to\infty} \mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f_n) = \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}}$. From this we conclude that *ii)* is not true.

i) $\Rightarrow$ *ii)*. Let us assume that *ii)* is not true. Then there exists an $\varepsilon_0 \in (0,\infty]$ such that for all $n \geq 1$ there exists a measurable function $f_n : X \to \mathbb{R}$ with $\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f_n) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} \geq \varepsilon_0$ and

$$\frac{1}{n} \geq \mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f_n) - \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} = \int_X \left| \mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\big(f_n(x)\big) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x} \right| d\mathrm{P}_X(x).$$

Hence there exists a sub-sequence (f_{n_i}) satisfying

$$\mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\big(f_{n_i}(x)\big) \to \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}$$

for P_X-almost all $x \in X$. Let us fix an $x \in X$ at which the convergence takes place and that additionally satisfies $\mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x} < \infty$, $\mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x} < \infty$, and $\delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) > 0$ for *all* $\varepsilon > 0$. For later use, note that the probability for such an element x is 1 since $\delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x)$ is monotonically increasing in ε. Now, for $\varepsilon > 0$, there exists an i_0 such that for all $i \geq i_0$ we have

$$\mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\big(f_{n_i}(x)\big) < \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x} + \delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x)\,.$$

By part *i*) of Lemma 3.14, this yields $\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}(f_{n_i}(x)) < \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x} + \varepsilon$, i.e., we have shown

$$\lim_{i\to\infty} \mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\big(f_{n_i}(x)\big) \;=\; \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\,. \tag{3.20}$$

Since the probability of the considered x was 1, the limit relation (3.20) holds for P_X-almost all $x \in X$, and hence we obtain $\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f_{n_i}) \to \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}}$ by Lebesgue's convergence theorem and (3.19). However, this contradicts the fact that $\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f_n) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} \geq \varepsilon_0$ holds for all $n \geq 1$. □

Theorem 3.17 shows that an almost surely strictly positive calibration function is *necessary* for a positive answer to Question 3.1, i.e., for having an implication of the form

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

for *all* sequences (f_n) of measurable functions. Moreover, Theorem 3.17 also shows that an almost surely strictly positive calibration function is *sufficient* for (3.21) if the *additional* assumption (3.19) holds. In this regard, note that in general this additional assumption is *not* superfluous. For details, we refer to Exercise 3.11.

Let us now recall that from a machine learning point of view we are not interested in a single distribution since we do not know the data-generating distribution P. However, we may know that P is a distribution of a certain type $\mathcal{Q}$, and consequently the following definition is of great importance in practical situations.

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

$$\delta_{\max}\left(\varepsilon, \mathrm{Q}, x\right) > 0\,.$$

Note that, using (3.15), we easily verify that L_{sur} is L_{tar}-calibrated with respect to $\mathcal{Q}$ if and only if for all $\varepsilon \in (0,\infty]$, $\mathrm{Q} \in \mathcal{Q}$, and $x \in X$ there is a $\delta \in (0,\infty]$ with

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

Now assume that our only information on the data-generating distribution P is that it is of some type $\mathcal{Q}$. Then Theorem 3.17 shows that we can only hope for a positive answer to Question 3.1 if our surrogate loss L_{sur} is L_{tar}-calibrated with respect to $\mathcal{Q}$. In this sense, calibration of L_{sur} is a first test on whether L_{sur} is a reasonable surrogate. The following corollary, whose proof is left as an exercise, shows that for some target losses this test is also sufficient.

Corollary 3.19. *Let X be a complete measurable space, $L_{\text{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\text{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be two losses, and $\mathcal{Q}$ be a set of distributions on Y. If L_{tar} is bounded, i.e., there is $B > 0$ with $L(x,y,t) \le B$ for all $(x,y,t) \in X \times Y \times \mathbb{R}$, then the following statements are equivalent:*

i) L_{sur} is L_{tar}-calibrated with respect to $\mathcal{Q}$.
ii) For all $\varepsilon \in (0,\infty]$ and all distributions P of type $\mathcal{Q}$ with $\mathcal{R}^*_{L_{\text{sur}},\mathrm{P}} < \infty$, there exists a $\delta \in (0,\infty]$ such that, for all measurable $f : X \to \mathbb{R}$, we have

$$\mathcal{R}_{L_{\text{sur}},\mathrm{P}}(f) < \mathcal{R}^*_{L_{\text{sur}},\mathrm{P}} + \delta \qquad \Longrightarrow \qquad \mathcal{R}_{L_{\text{tar}},\mathrm{P}}(f) < \mathcal{R}^*_{L_{\text{tar}},\mathrm{P}} + \varepsilon\,.$$

Recall that both the classification loss and the density level detection loss are bounded losses, and consequently the preceding corollary applies to these target losses. Moreover, for the classification loss being the target loss and the least squares or the hinge loss being the surrogate loss, we have already shown in Example 3.16 that the corresponding calibration function is strictly positive. Consequently, Corollary 3.19 shows that both loss functions are reasonable surrogates in an asymptotic sense. However, we have already seen in Zhang's inequality, see Theorem 2.31, that there is even a strong *quantitative* relationship between the excess classification risk and the excess hinge risk. Such stronger relationships are studied in the next section in more detail.

3.3 Inequalities between Excess Risks

If one wants to find a good surrogate loss L_{sur} for a given target loss L_{tar}, then implication (3.18) is in some sense a minimal requirement. However, we have already indicated in Question 3.2 that in many cases one actually needs quantified versions of (3.18), e.g., in terms of *inequalities* between the excess L_{tar}-risk and the excess L_{sur}-risk. Considering Theorem 3.17, such inequalities are readily available if, for all $\varepsilon > 0$, we *know* a $\delta(\varepsilon) > 0$ such that implication (3.18) holds for all measurable $f : X \to \mathbb{R}$. Indeed, for f with $\varepsilon := \mathcal{R}_{L_{\text{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{tar}},\mathrm{P}} > 0$, we have $\mathcal{R}_{L_{\text{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\mathrm{P}} \ge \delta(\varepsilon)$, or in other words

$$\delta\bigl(\mathcal{R}_{L_{\text{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{tar}},\mathrm{P}}\bigr) \;\le\; \mathcal{R}_{L_{\text{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\mathrm{P}}\,. \tag{3.23}$$

In addition, if we define $\delta(0) := 0$, then this inequality actually holds for *all* measurable $f : X \to \mathbb{R}$. Unfortunately, however, the proof of Theorem 3.17 does not provide a constructive way to find a value for $\delta(\varepsilon)$, and hence we have so far no method to establish inequalities of the form (3.23). This problem is resolved in the following theorems for which we first introduce the Fenchel-Legendre bi-conjugate of a function.

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 $I = [0,\infty)$.*

Note that if $g : [0,B] \to [0,\infty)$ is a strictly positive and increasing function on $(0,B]$ with $g(0) = 0$, then Lemma A.6.20 shows that its bi-conjugate g^{**} is also strictly positive on $(0,B]$. Furthermore, a similar result holds for continuous functions (see Lemma A.6.21). However, these results are in general false if one considers functions on $I := [0,\infty)$, as, e.g., the square root $\sqrt{\cdot} : [0,\infty) \to [0,\infty)$ shows.

Besides the Fenchel-Legendre bi-conjugate, we also need some additional notations and definitions. To this end, let X be a complete measurable space, $L_{\text{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss function, and P be a distribution on $X \times Y$ such that $\mathcal{R}^*_{L_{\text{tar}},\mathrm{P}} < \infty$. For a measurable function $f : X \to \mathbb{R}$, we write

$$B_f := \left\| x \mapsto \left(\mathcal{C}_{L_{\text{tar}},\mathrm{P}(\,\cdot\,|x),x}(f(x)) - \mathcal{C}^*_{L_{\text{tar}},\mathrm{P}(\,\cdot\,|x),x} \right) \right\|_\infty , \tag{3.24}$$

i.e., B_f is the supremum of the excess inner target risk with respect to f. Note that in the following considerations we do *not* require $B_f < \infty$.

Our first two results on inequalities between excess risks will only assume that the involved distribution P is of some type $\mathcal{Q}$. In this case, the following notion of calibration will be crucial.

Definition 3.21. *Let $L_{\text{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\text{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be two losses and $\mathcal{Q}$ be a set of distributions on Y. Then the* ***uniform calibration function*** *with respect to $\mathcal{Q}$ is defined by*

$$\delta_{\max}(\varepsilon, \mathcal{Q}) := \inf_{\substack{\mathrm{Q}\in\mathcal{Q} \\ x\in X}} \delta_{\max}\left(\varepsilon, \mathrm{Q}, x\right) , \qquad \varepsilon \in [0,\infty].$$

Moreover, we say that L_{sur} is ***uniformly L_{tar}-calibrated*** *with respect to $\mathcal{Q}$ if $\delta_{\max}(\varepsilon,\mathcal{Q}) > 0$ for all $\varepsilon \in (0,\infty]$.*

Obviously, every uniformly calibrated loss function is calibrated; however, the converse implication does not hold in general. Since we will see important examples of the latter statement in Section 3.7, we do not present such an example here. Finally, note that an alternative definition of $\delta_{\max}(\varepsilon,\mathcal{Q})$ can be found in Exercise 3.5.

Now we are well-prepared to formulate our first result, which establishes inequalities between excess risks of different loss functions.

Theorem 3.22 (Uniform calibration inequalities). *Let X be a complete measurable space, $L_{\mathrm{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be losses, and $\mathcal{Q}$ be a set of distributions on Y. Moreover, let $\delta : [0,\infty] \to [0,\infty]$ be an increasing function such that*

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

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

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

where $\delta^{**}_{B_f} : [0, B_f] \to [0,\infty]$ is the Fenchel-Legendre biconjugate of $\delta_{|[0,B_f]}$ and B_f is defined by (3.24).

Proof. Inequalities (3.16) and (3.25) together with $\mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} < \infty$ and $\mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} < \infty$ give

$$\delta\big(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\big) \;\leq\; \mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x} \tag{3.27}$$

for P_X-almost all $x \in X$ and all $t \in \mathbb{R}$. For a measurable function $f : X \to \mathbb{R}$ with $\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) < \infty$, Jensen's inequality together with the definition of B_f, $\delta^{**}_{B_f}(\,\cdot\,) \leq \delta(\,\cdot\,)$, and (3.27) now yields

$$\begin{aligned}
&\delta^{**}_{B_f}\big(\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}}\big) \\
&\leq \int_X \delta^{**}_{B_f}\Big(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\Big) d\mathrm{P}_X(x) \\
&\leq \int_X \mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\, d\mathrm{P}_X(x) \\
&= \mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}}\,.
\end{aligned}$$

Finally, for $f : X \to \mathbb{R}$ with $\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) = \infty$, we have $B_f = \infty$. If $\delta^{**}_\infty(\infty) = 0$, there is nothing to prove, and hence let us assume $\delta^{**}_\infty(\infty) > 0$. Then (3.25) implies $\delta(0) = 0$ and hence δ^{**}_∞ is increasing because of its convexity and $\delta^{**}_\infty(0) = \delta(0) = 0$. Consequently, if δ^{**}_∞ is finite on $[0,\infty)$, then there exists a $t_0 \geq 0$ and a $c_0 > 0$ such that the (Lebesgue)-almost surely defined derivative of δ^{**}_∞ satisfies $(\delta^{**}_\infty)'(t) \geq c_0$ for almost all $t \geq t_0$. By Lebesgue's version of the fundamental theorem of calculus, see Theorem A.6.6, we then find constants $c_1, c_2 \in (0,\infty)$ with $t \leq c_1\delta^{**}_\infty(t) + c_2$ for all $t \in [0,\infty]$. On the other hand, if there is a $t_0 > 0$ with $\delta^{**}_\infty(t_0) = \infty$, we have $t \leq c_1\delta^{**}_\infty(t) + c_2$ for $c_1 := 1$, $c_2 := t_0$, and all $t \in [0,\infty]$. In both cases, (3.27) now yields

$$\begin{aligned}
\infty &= \int_X \Big(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\Big) d\mathrm{P}_X(x) \\
&\leq c_1 \int_X \delta^{**}_\infty\Big(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\Big) d\mathrm{P}_X(x) + c_2 \\
&\leq c_1\big(\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}}\big) + c_2
\end{aligned}$$

and hence we have $\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} = \infty$. □

Note that if L_{sur} is uniformly L_{tar}-calibrated with respect to $\mathcal{Q}$ and the function $f : X \to \mathbb{R}$ satisfies $B_f < \infty$, then Lemma A.6.20 shows that the bi-conjugate of $\delta_{\max}(\,\cdot\,, \mathcal{Q})_{|[0,B_f]}$ is strictly positive on $(0, B_f]$. Consequently, Theorem 3.22 gives a *non-trivial* inequality in this case.

Let us now illustrate the theory we have developed so far by a simple example dealing with the least squares and the hinge loss.

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

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

for all $\varepsilon > 0$. For the least squares loss, this yields

$$\delta_{\max,L_{\text{class}},L}(\varepsilon, \mathcal{Q}_Y) = \varepsilon^2\,, \qquad\qquad \varepsilon > 0,$$

which by Theorem 3.22 implies that, for all measurable $f : X \to \mathbb{R}$, we have

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

On the other hand, for the hinge loss, we find $\delta_{\max,L_{\text{class}},L}(\varepsilon, \mathcal{Q}_Y) = \varepsilon$ for all $\varepsilon > 0$, and hence Theorem 3.22 recovers Zhang's inequality. ◁

The following result shows that uniform calibration is also *necessary* to establish non-trivial inequalities that hold for *all* distributions of some type.

Theorem 3.24. *Let X be a complete measurable space, $L_{\text{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\text{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be two losses, and $\mathcal{Q}$ be a set of distributions on Y such that $\mathcal{C}^*_{L_{\text{tar}},\mathrm{Q},x} < \infty$ and $\mathcal{C}^*_{L_{\text{sur}},\mathrm{Q},x} < \infty$ for all $x \in X$ and $\mathrm{Q} \in \mathcal{Q}$. Furthermore, let $\delta : [0,\infty] \to [0,\infty]$ be increasing with $\delta(0) = 0$ and $\delta(\varepsilon) > 0$ for all $\varepsilon > 0$. If for all distributions P of type $\mathcal{Q}$ satisfying $\mathcal{R}^*_{L_{\text{tar}},\mathrm{P}} < \infty$ and $\mathcal{R}^*_{L_{\text{sur}},\mathrm{P}} < \infty$ and all measurable $f : X \to \mathbb{R}$ we have*

$$\delta\big(\mathcal{R}_{L_{\text{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{tar}},\mathrm{P}}\big) \;\leq\; \mathcal{R}_{L_{\text{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\mathrm{P}}\,,$$

then L_{sur} is uniformly L_{tar}-calibrated with respect to $\mathcal{Q}$.

Proof. Let us fix an $x \in X$ and a $\mathrm{Q} \in \mathcal{Q}$. Furthermore, let P be the distribution on $X \times Y$ with $\mathrm{P}_X = \delta_{\{x\}}$ and $\mathrm{P}(\,\cdot\,|x) = \mathrm{Q}$. Then P is of type $\mathcal{Q}$, and we have both $\mathcal{R}_{L_i,\mathrm{P}}(f) = \mathcal{C}_{L_i,\mathrm{Q},x}(f(x))$ and $\mathcal{R}^*_{L_i,\mathrm{P}} = \mathcal{C}^*_{L_i,\mathrm{Q},x} < \infty$ for $i = \{\text{tar},\text{sur}\}$ and all measurable $f : X \to \mathbb{R}$. Consequently, our assumption yields

$$\delta\big(\mathcal{C}_{L_{\text{tar}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\text{tar}},\mathrm{Q},x}\big) \;\leq\; \mathcal{C}_{L_{\text{sur}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\text{sur}},\mathrm{Q},x}\,, \qquad\qquad t \in \mathbb{R}.$$

Now let $\varepsilon > 0$ and $t \in \mathcal{M}_{L_{\text{sur}},\mathrm{Q},x}(\delta(\varepsilon))$. Then we have $\mathcal{C}_{L_{\text{sur}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\text{sur}},\mathrm{Q},x} < \delta(\varepsilon)$, and hence the inequality above yields $\delta(\mathcal{C}_{L_{\text{tar}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\text{tar}},\mathrm{Q},x}) < \delta(\varepsilon)$. Since δ is monotonically increasing, the latter shows $\mathcal{C}_{L_{\text{tar}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\text{tar}},\mathrm{Q},x} < \varepsilon$, i.e., we have found $\mathcal{M}_{L_{\text{sur}},\mathrm{Q},x}(\delta(\varepsilon)) \subset \mathcal{M}_{L_{\text{tar}},\mathrm{Q},x}(\varepsilon)$. Equation (3.15) then shows $\delta(\varepsilon) \leq \delta_{\max}(\varepsilon, \mathrm{Q}, x)$, and hence L_{sur} is uniformly L_{tar}-calibrated with respect to $\mathcal{Q}$. □

It will turn out in Sections 3.7 and 3.9, for example, that in many situations we have calibrated losses that are not uniformly calibrated. We have just seen that in such cases we need assumptions on P stronger than the $\mathcal{Q}$-type to establish inequalities. The following theorem presents a result in this direction.

Theorem 3.25 (General calibration inequalities). *Let X be a complete measurable space, $L_{\rm tar} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\rm sur} : X \times Y \times \mathbb{R} \to [0,\infty)$ be two losses, and P be a distribution on $X \times Y$ such that $\mathcal{R}^*_{L_{\rm tar},\mathrm{P}} < \infty$ and $\mathcal{R}^*_{L_{\rm sur},\mathrm{P}} < \infty$. Assume that there exist a $p \in (0,\infty]$ and functions $b : X \to [0,\infty]$ and $\delta : [0,\infty) \to [0,\infty)$ such that*

$$\delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) \;\geq\; b(x)\,\delta(\varepsilon)\,, \qquad\qquad \varepsilon \geq 0,\, x \in X, \tag{3.28}$$

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

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

where $\bar{\delta}^{**}_{B_f} : [0, B_f] \to [0,\infty]$ is the Fenchel-Legendre biconjugate of $\bar{\delta}_{|[0,B_f]}$ and B_f is defined by (3.24).

Proof. Let us first consider the case $\mathcal{R}_{L_{\rm tar},\mathrm{P}}(f) < \infty$. Since $\bar{\delta}^{**}_{B_f}$ is convex and satisfies $\bar{\delta}^{**}_{B_f}(\varepsilon) \leq \bar{\delta}(\varepsilon)$ for all $\varepsilon \in [0, B_f]$, we see by Jensen's inequality that

$$\bar{\delta}^{**}_{B_f}\big(\mathcal{R}_{L_{\rm tar},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\rm tar},\mathrm{P}}\big) \;\leq\; \int_X \bar{\delta}\big(\mathcal{C}_{L_{\rm tar},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\rm tar},\mathrm{P}(\,\cdot\,|x),x}\big)\,d\mathrm{P}_X(x)\,.$$

Moreover, using (3.28) and (3.16), we obtain

$$b(x)\,\delta\big(\mathcal{C}_{L_{\rm tar},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\rm tar},\mathrm{P}(\,\cdot\,|x),x}\big) \;\leq\; \mathcal{C}_{L_{\rm sur},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\rm sur},\mathrm{P}(\,\cdot\,|x),x}$$

for P_X-almost all $x \in X$ and all $t \in \mathbb{R}$. Now note that for $q := (1+p)/p$ the conjugate exponent satisfies $q' = 1 + p = pq$. By the definition of $\bar{\delta}$ and Hölder's inequality in the form of $\mathbb{E}|hg|^{1/q} \leq (\mathbb{E}|h|^{q'/q})^{1/q'}(\mathbb{E}|g|)^{1/q}$, we thus find

$$\begin{aligned}
&\int_X \bar{\delta}\big(\mathcal{C}_{L_{\rm tar},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\rm tar},\mathrm{P}(\,\cdot\,|x),x}\big)\,d\mathrm{P}_X(x)\\
&\leq \int_X \big(b(x)\big)^{-\frac{1}{q}}\Big(\mathcal{C}_{L_{\rm sur},\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big) - \mathcal{C}^*_{L_{\rm sur},\mathrm{P}(\,\cdot\,|x),x}\Big)^{\frac{1}{q}}\,d\mathrm{P}_X(x)\\
&\leq \left(\int_X b^{-p} d\mathrm{P}_X\right)^{\frac{1}{qp}}\left(\int_X \mathcal{C}_{L_{\rm sur},\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big) - \mathcal{C}^*_{L_{\rm sur},\mathrm{P}(\,\cdot\,|x),x}\,d\mathrm{P}_X(x)\right)^{1/q}\\
&= \|b^{-1}\|^{\frac{1}{q}}_{L_p(\mathrm{P}_X)}\big(\mathcal{R}_{L_{\rm sur},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\rm tar},\mathrm{P}}\big)^{1/q}\,.
\end{aligned}$$

Combining this estimate with our first estimate then gives the assertion in the case $\mathcal{R}_{L_{\rm tar},\mathrm{P}}(f) < \infty$. On the other hand, if $\mathcal{R}_{L_{\rm tar},\mathrm{P}}(f) = \infty$, we have

$B_f = \infty$. If $\bar{\delta}^{**}_{\infty}(\infty) = 0$, there is nothing to prove and hence we restrict our considerations to the case where $\bar{\delta}^{**}_{\infty}(\infty) > 0$. In this case, the proof of Theorem 3.22 has already shown that then there exist constants $c_1, c_2 \in (0, \infty)$ such that $t \leq c_1 \bar{\delta}^{**}_{\infty}(t) + c_2$ for all $t \in [0, \infty]$. From this we obtain

$$\begin{aligned}
\infty &= \mathcal{R}_{L_{\text{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{tar}},\mathrm{P}} \\
&\leq c_1 \int_X \bar{\delta}^{**}_{\infty}\big(\mathcal{C}_{L_{\text{tar}},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\text{tar}},\mathrm{P}(\,\cdot\,|x),x}\big)\, d\mathrm{P}_X(x) + c_2 \\
&\leq c_1 \int_X \big(b(x)\big)^{-\frac{1}{q}} \Big(\mathcal{C}_{L_{\text{sur}},\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big) - \mathcal{C}^*_{L_{\text{sur}},\mathrm{P}(\,\cdot\,|x),x}\Big)^{\frac{1}{q}} d\mathrm{P}_X(x) + c_2\,,
\end{aligned}$$

where the last step is analogous to our considerations in the case $\mathcal{R}_{L_{\text{tar}},\mathrm{P}}(f) < \infty$. Using $b^{-1} \in L_p(\mathrm{P}_X)$ and Hölder's inequality, we then conclude from the estimate above that $\mathcal{R}_{L_{\text{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\mathrm{P}} = \infty$. □

The condition $b^{-1} \in L_p(\mathrm{P}_X)$ in the preceding theorem measures how much the calibration function $\delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x)$ violates a uniform lower bound of the form $\delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) \geq c\,\delta(\varepsilon)$, $\varepsilon \in [0, \infty]$. Indeed, the larger we can choose p in condition (3.28), the more the function b is away from the critical level 0, and thus the closer condition (3.28) is to a uniform lower bound. In the extremal case $p = \infty$, condition (3.28) actually becomes a uniform bound, and the inequality of Theorem 3.25 equals the inequality of Theorem 3.22. Finally, for $\delta(\varepsilon) := \varepsilon^r$, $\varepsilon \geq 0$, the function $\bar{\delta}(\varepsilon) := \delta^{\frac{p}{p+1}}(\varepsilon) = \varepsilon^{\frac{rp}{p+1}}$ is convex if $r \geq 1 + 1/p$. In this case, we can thus omit the Fenchel-Legendre biconjugate in Theorem 3.25 and obtain the simpler inequality

$$\mathcal{R}_{L_{\text{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{tar}},\mathrm{P}} \;\leq\; \|b^{-1}\|^{1/r}_{L_p(\mathrm{P}_X)} \big(\mathcal{R}_{L_{\text{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\mathrm{P}}\big)^{1/r}\,.$$

Here, the condition $r \geq 1 + 1/p$ means that we have to increase the convexity of δ if we wish to weaken the uniformity of the calibration.

Our last goal in this section is to improve the inequalities above for the following type of loss, which will be of great utility in the next sections.

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** with respect to (A, h) if*

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

Every detection loss function is obviously measurable and hence an unsupervised loss function. In addition, for $x \in X$ and $t \in \mathbb{R}$, we have

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

Since detection losses will play an important role for *both* supervised and unsupervised learning scenarios let us now establish some specific results for this class of target loss function. We begin with the following theorem, whose proof is similar to the proof of Corollary 3.19 and hence is left as Exercise 3.7.

Theorem 3.27 (Asymptotic calibration for detection losses). *Let X be a complete measurable space and $L_{\mathrm{tar}} : X \times \mathbb{R} \to [0,\infty)$ be a detection loss with respect to some (A,h). Moreover, 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 following statements are equivalent:*

i) L_{sur} is L_{tar}-calibrated with respect to $\mathcal{Q}$.

ii) For all distributions P of type $\mathcal{Q}$ that satisfy $h \in L_1(\mathrm{P}_X)$ and $\mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} < \infty$ and all $\varepsilon \in (0,\infty]$, there exists a $\delta \in (0,\infty]$ such that for all measurable $f : X \to \mathbb{R}$ we have

$$\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) < \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} + \delta \qquad \Longrightarrow \qquad \mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) < \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} + \varepsilon\,.$$

If the target loss is a detection loss, then we can, of course, establish calibration inequalities by Theorems 3.22 and 3.25. However, using the specific form of detection losses, one can often improve the resulting inequalities, as we will discuss after the following rather general theorem.

Theorem 3.28 (Calibration inequalities for detection losses). *Let X be a complete measurable space, $L_{\mathrm{tar}} : X \times \mathbb{R} \to [0,\infty)$ be a detection loss with respect to (A,h), $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, and P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} < \infty$ and $\mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} < \infty$. For $s > 0$, we write*

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

If there exist constants $c > 0$ and $\alpha \in (0,\infty]$ such that

$$\int_X \mathbf{1}_{B(s)} h \, d\mathrm{P}_X \;\leq\; (c\,s)^{\alpha}\,, \qquad\qquad s > 0, \tag{3.30}$$

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

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

Proof. We write $\mathcal{C}_{\mathrm{tar},x}(f) := \mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}(f(x)) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}$ for $x \in X$ and measurable $f : X \to \mathbb{R}$. Furthermore, for $s > 0$, we write

$$C(s) := \big\{ x \in X : A(x) \neq \mathbb{R}, \text{ and } \delta_{\max}(h(x), \mathrm{P}(\,\cdot\,|x), x) \geq s\,h(x) \big\}\,.$$

By (3.16) and (3.29), we then obtain

$$\begin{aligned}
&\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} \\
&= \int_{B(s)} \mathbf{1}_A(x, f(x))\, h(x)\, d\mathrm{P}_X(x) + \int_{C(s)} \mathbf{1}_A(x, f(x))\, h(x)\, d\mathrm{P}_X(x) \\
&\leq \int_X \mathbf{1}_{B(s)} h \, d\mathrm{P}_X + s^{-1} \int_{C(s)} \delta_{\max}\big(h(x), \mathrm{P}(\,\cdot\,|x), x\big)\, \mathbf{1}_A(x, f(x))\, d\mathrm{P}_X(x) \\
&\leq (c\,s)^{\alpha} + s^{-1} \int_{C(s)} \delta_{\max}\big(\mathcal{C}_{\mathrm{tar},x}(f), \mathrm{P}(\,\cdot\,|x), x\big)\, d\mathrm{P}_X(x) \\
&\leq (c\,s)^{\alpha} + s^{-1} \big(\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}}\big)\,.
\end{aligned}$$

If $\alpha < \infty$, we now choose $s := (\alpha c^\alpha)^{-\frac{1}{\alpha+1}} (\mathcal{R}_{L_{\text{sur}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\text{P}})^{\frac{1}{\alpha+1}}$. Using $\alpha^{-\frac{\alpha}{\alpha+1}} + \alpha^{\frac{1}{\alpha+1}} \leq 2$ then yields the assertion. Furthermore, for $\alpha = \infty$, the assertion follows by setting $s^{-1} := 2c$. □

The preceding theorem can improve the inequalities we obtained for general target losses in various cases. The following two remarks illustrate this.

Remark 3.29. For detection losses with $h = \mathbf{1}_X$, Theorem 3.28 yields an improvement over Theorem 3.25. Indeed, if (3.28) is satisfied for $\delta(\varepsilon) = \varepsilon^q$ and a $b : X \to [0, \infty]$ with $b^{-1} \in L_p(\text{P}_X)$ and $q \geq \frac{p+1}{p}$, then Theorem 3.25 gives

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

On the other hand, some calculations show $B(s) \subset \{x \in X : b(x) < s\}$, and since $b^{-1} \in L_p(\text{P}_X)$ implies

$$\text{P}_X\big(\{x \in X : b(x) < s\}\big) \leq \|b^{-1}\|^p_p \, s^p\,, \qquad\qquad s > 0,$$

we find (3.30) for $c := \|b^{-1}\|_{L_p(\text{P}_X)}$ and $\alpha := p$. Theorem 3.28 thus yields

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

Now note that for $q > \frac{p+1}{p}$, (3.32) is sharper than (3.31) whenever the excess risk $\mathcal{R}_{L_{\text{sur}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\text{P}}$ is sufficiently small. ◁

Remark 3.30. In some cases, Theorem 3.28 also improves the inequalities of Theorem 3.22. Indeed, if L_{sur} is uniformly L_{tar}-calibrated with respect to some class $\mathcal{Q}$ of distributions and the uniform calibration function satisfies $\delta_{\max}(\,\cdot\,, \mathcal{Q}) \geq c_q \varepsilon^q$ for some $q > 1$, $c_q > 0$, and all $\varepsilon \geq 0$, then Theorem 3.22 gives

$$\mathcal{R}_{L_{\text{tar}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{tar}},\text{P}} \;\leq\; c_q^{-1/q} \big(\mathcal{R}_{L_{\text{sur}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\text{P}}\big)^{1/q} \tag{3.33}$$

for all measurable functions $f : X \to \mathbb{R}$. However, an easy calculation shows that the assumptions above imply $B(s) \subset \{x \in X : 0 < h(x) < (s/c_q)^{1/(q-1)}\}$. Consequently, if we have constants $C > 0$ and $\beta \in (0, \infty]$ such that

$$\text{P}_X\big(\big\{x \in X : 0 < h(x) < s\big\}\big) \;\leq\; (C\,s)^\beta\,, \qquad\qquad s > 0, \tag{3.34}$$

then it is easy to check that (3.30) is satisfied for $c = c_q^{-1} C^{\frac{\beta q - \beta}{\beta+1}}$ and $\alpha := \frac{\beta+1}{q-1}$. Theorem 3.28 thus yields

$$\mathcal{R}_{L_{\text{tar}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{tar}},\text{P}} \;\leq\; 2\, c_q^{-\frac{\beta+1}{\beta+q}} C^{\frac{\beta q - \beta}{\beta+q}} \left(\mathcal{R}_{L_{\text{sur}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\text{P}}\right)^{\frac{\beta+1}{\beta+q}}. \tag{3.35}$$

Now note that for $q > 1$, we have $\frac{\beta+1}{\beta+q} > \frac{1}{q}$, and thus (3.35) is sharper than (3.33) whenever $\mathcal{R}_{L_{\text{sur}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\text{P}}$ is sufficiently small. ◁

3.4 Surrogates for Unweighted Binary Classification

In this section, we apply the general theory on surrogate loss functions developed in the previous sections to the standard binary classification scenario. The result of this section will be important for Section 8.5, where we investigate SVMs for classification that do not use the hinge loss as a surrogate.

Let us first recall (see Example 2.4) that in binary classification we consider the label space $Y := \{-1, 1\}$ together with the supervised loss L_{class}. In the following, we write $\mathcal{Q}_Y$ for the set of all distributions on Y. Moreover, recall that any distribution $\mathrm{Q} \in \mathcal{Q}_Y$ can be uniquely described by an $\eta \in [0,1]$ using the identification $\eta = \mathrm{Q}(\{1\})$. If $L : Y \times \mathbb{R} \to [0,\infty)$ is a supervised loss, we therefore use the notation

$$\begin{aligned} \mathcal{C}_{L,\eta}(t) &:= \mathcal{C}_{L,\mathrm{Q}}(t)\,, \qquad & t \in \mathbb{R}, \\ \mathcal{C}^*_{L,\eta} &:= \mathcal{C}^*_{L,\mathrm{Q}}\,, \end{aligned} \tag{3.36}$$

as well as $\mathcal{M}_{L,\eta}(0^+) := \mathcal{M}_{L,\mathrm{Q}}(0^+)$, $\mathcal{M}_{L,\eta}(\varepsilon) := \mathcal{M}_{L,\mathrm{Q}}(\varepsilon)$, and $\delta_{\max}(\varepsilon,\eta) := \delta_{\max}(\varepsilon,\mathrm{Q})$ for $\varepsilon \in [0,\infty]$. Note that, by the special structure of margin-based losses and the distributions $\mathrm{Q} \in \mathcal{Q}_Y$, we have the following symmetries:

$$\begin{aligned} \mathcal{C}_{L,\eta}(t) &= \mathcal{C}_{L,1-\eta}(-t) \qquad && \text{and} \qquad && \mathcal{C}^*_{L,\eta} = \mathcal{C}^*_{L,1-\eta}\,, \\ \mathcal{M}_{L,\eta}(\varepsilon) &= -\mathcal{M}_{L,1-\eta}(\varepsilon) \qquad && \text{and} \qquad && \mathcal{M}_{L,\eta}(0^+) = -\mathcal{M}_{L,1-\eta}(0^+)\,. \end{aligned}$$

Furthermore, it is interesting to note that the quantity $2\eta-1$, which will occur at many places in the following results, is the *expectation* of the corresponding Q, i.e., $\mathbb{E}\mathrm{Q} := \mathbb{E}_{\mathrm{Q}}\,\mathrm{id}_Y = 2\eta-1$. Before we present our first results, let us finally simplify our nomenclature.

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_{class}-calibrated with respect to $\mathcal{Q}_Y$.*

Now our first aim is to compute the calibration function $\delta_{\max,L_{\text{class}},L}(\,\cdot\,,\eta)$ for supervised surrogates L of L_{class}.

Lemma 3.32 (Calibration function). *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss. Then, for all $\eta \in [0,1]$ and $\varepsilon \in (0,\infty]$, we have*

$$\delta_{\max,L_{\text{class}},L}(\varepsilon,\eta) = \begin{cases} \infty & \text{if } \varepsilon > |2\eta-1| \\ \inf_{t\in\mathbb{R}:(2\eta-1)\operatorname{sign} t \le 0} \bigl(\mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta}\bigr) & \text{if } \varepsilon \le |2\eta-1|. \end{cases}$$

Proof. The assertion immediately follows from the formula

$$\mathcal{M}_{L_{\text{class}},\eta}(\varepsilon) = \begin{cases} \mathbb{R} & \text{if } \varepsilon > |2\eta-1| \\ \{t \in \mathbb{R} : (2\eta-1)\operatorname{sign} t > 0\} & \text{if } 0 < \varepsilon \le |2\eta-1|\,, \end{cases}$$

which we derived in Example 3.8, and $\inf\emptyset = \infty$. □

The formula for the calibration function presented in Lemma 3.32 implies that $\delta_{\max}(\,\cdot\,,\eta)$ is a step function that only attains one value different from 0 and ∞. This particular form of the calibration function is the key ingredient of the following considerations on the relation between classification calibration and uniform classification calibration. We begin with a preliminary lemma.

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*

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

Then the following statements are true:

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

Proof. i). Let us first assume that L is classification calibrated. We fix an $\eta \neq 1/2$. Then Lemma 3.32 together with $\operatorname{sign} 0 = 1$ shows $\mathcal{C}_{L,\eta}(0) > \mathcal{C}^*_{L,\eta}$ if $\eta \in [0,1/2)$. Moreover, if $\eta \in (1/2,1]$, we find the same inequality by

$$\mathcal{C}_{L,\eta}(0) - \mathcal{C}^*_{L,\eta} = \mathcal{C}_{L,1-\eta}(0) - \mathcal{C}^*_{L,1-\eta} > 0\,.$$

Finally, Lemma 3.32 yields

$$\inf_{t\in\mathbb{R}:(2\eta-1)t<0} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} \;\ge\; \delta_{\max}(\varepsilon,\eta) \;>\; 0 \tag{3.38}$$

for $0 < \varepsilon \le |2\eta - 1|$, and hence we find $H(\eta) > 0$. Conversely, Lemma 3.32 gives $\delta_{\max}(\varepsilon,\eta) \ge H(\eta)$ for all $0 < \varepsilon \le |2\eta - 1|$, and hence L is classification calibrated if $H(\eta) > 0$ for all $\eta \neq 1/2$.

ii). Since there is nothing to prove in the case $\eta = 1/2$, we assume $\eta \neq 1/2$. Now, if L is continuous, then $\mathcal{C}_{L,\eta}(\,\cdot\,)$ is continuous at 0, and hence we have

$$\delta_{\max}(\varepsilon,\eta) \le \inf_{t\in\mathbb{R}:(2\eta-1)t<0} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} = \inf_{t\in\mathbb{R}:(2\eta-1)t\le 0} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} = H(\eta)$$

by (3.38). Moreover, for $0 < \varepsilon \le |2\eta - 1|$, we always have $\delta_{\max}(\varepsilon,\eta) \ge H(\eta)$.

iii). The equality $H(1/2) = 0$ is trivial, and $H(\eta) = H(1-\eta)$, $\eta \in [0,1]$, immediately follows from symmetries mentioned at the beginning of this section. In order to prove the continuity of H, we now define

$$\begin{aligned} h(\eta) &= \inf_{t\in\mathbb{R}:(2\eta-1)t\le 0} \mathcal{C}_{L,\eta}(t)\,, \\ g^+(\eta) &= \inf_{t\le 0} \mathcal{C}_{L,\eta}(t)\,, \\ g^-(\eta) &= \inf_{t\ge 0} \mathcal{C}_{L,\eta}(t)\,, \end{aligned}$$

for $\eta \in [0,1]$. Then the functions $g^+ : [0,1] \to [0,\infty)$ and $g^- : [0,1] \to [0,\infty)$ can be defined by suprema taken over affine linear functions in $\eta \in$

$\mathbb{R}$, and since g^+ and g^- are also finite for $\eta \in [0,1]$, we find by Lemma A.6.4 that g^+ and g^- are continuous at every $\eta \in [0,1]$. Moreover, we have $\mathcal{C}^*_{L,\eta} = \min\{g^+(\eta), g^-(\eta)\}$ for all $\eta \in [0,1]$, and hence $\eta \mapsto \mathcal{C}^*_{L,\eta}$ is continuous. Finally, we have $h(\eta) = g^-(\eta)$ for $\eta \in [0,1/2)$, $h(\eta) = g^+(\eta)$ for $\eta \in (1/2,1]$, and $h(1/2) = \min\{g^+(1/2), g^-(1/2)\} = g^-(1/2) = g^+(1/2)$. This shows that $h : [0,1] \to [0,\infty)$ is continuous, and by combining these results we then obtain the continuity of H. □

Now we can establish the main result of this section, which shows that classification calibrated, margin-based losses are *uniformly* classification calibrated. In addition, it provides a lower bound of the Fenchel-Legendre biconjugate (see Definition 3.20) of the uniform calibration function $\delta_{\max}(\cdot, \mathcal{Q}_Y)$.

Theorem 3.34 (Classification calibration). *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a margin-based loss. Then the following statements 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

$$\delta(\varepsilon) := H\Big(\frac{1+\varepsilon}{2}\Big), \qquad \varepsilon \in [0,1],$$

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

$$\delta^{**}(\varepsilon) \le \delta^{**}_{\max, L_{\text{class}}, L}(\varepsilon, \mathcal{Q}_Y), \qquad \varepsilon \in [0,1], \tag{3.39}$$

and both quantities are actually equal if L is continuous. Finally, if L is classification calibrated, we have $\delta^{**}(\varepsilon) > 0$ for all $\varepsilon \in (0,1]$.

Proof. We begin with a preliminary consideration. To this end, let us fix an $\varepsilon \in (0,1]$. Then, by Lemma 3.32 and the symmetry of H around $1/2$, we find

$$\delta_{\max}(\varepsilon, \mathcal{Q}_Y) = \inf_{|2\eta-1| \ge \varepsilon} \delta_{\max}(\varepsilon, \eta) \ge \inf_{|2\eta-1| \ge \varepsilon} H(\eta) = \inf_{\eta \ge \frac{\varepsilon+1}{2}} H(\eta) =: \tilde{\delta}(\varepsilon),$$

and with $\tilde{\delta}(0) := 0$ we also have $\delta_{\max}(0, \mathcal{Q}_Y) = \tilde{\delta}(0)$.

i) $\Leftrightarrow$ *ii)*. Since *ii)* $\Rightarrow$ *i)* is trivial, it suffices to show *i)* $\Rightarrow$ *ii)*. To this end, recall that H is continuous and strictly positive on all intervals $[\frac{\varepsilon+1}{2}, 1]$, $\varepsilon \in (0,1]$, by Lemma 3.33, and consequently we have $\tilde{\delta}(\varepsilon) > 0$ for all $\varepsilon > 0$. From this we find $\delta_{\max}(\varepsilon, \mathcal{Q}_Y) > 0$ for all $\varepsilon > 0$ by our preliminary consideration.

In order to show (3.39), recall that $\tilde{\delta}(\varepsilon) \le \delta_{\max}(\varepsilon, \mathcal{Q}_Y)$ holds for all $\varepsilon \in [0,1]$, and hence we find $\tilde{\delta}^{**}(\varepsilon) \le \delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y)$ for all $\varepsilon \in [0,1]$. Furthermore, we obviously have $\tilde{\delta}(\varepsilon) = \inf_{\varepsilon' \ge \varepsilon} \delta(\varepsilon')$, and hence Lemma A.6.21 gives $\delta^{**} = \tilde{\delta}^{**}$. In addition, if L is continuous, then our preliminary consideration together with Lemma 3.33 actually yields $\tilde{\delta}(\varepsilon) = \delta_{\max}(\varepsilon, \mathcal{Q}_Y)$ for all $\varepsilon \in [0,1]$. Repeating the arguments above thus shows $\delta^{**}(\varepsilon) = \delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y)$ for all $\varepsilon \in [0,1]$.

Table 3.1. Some common margin-based losses and the corresponding values for $H(\eta)$, $\eta \in [0,1]$, and $\delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y)$, $\varepsilon \in [0,1]$. All results easily follow from Theorem 3.36. For the logistic loss, we used the abbreviation $\Lambda(x) := x\ln(x)$. Note that if one wants to derive inequalities for the logistic loss using the above form of $\delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y)$, it is useful to know that $\varepsilon^2 \le \Lambda(1+\varepsilon) + \Lambda(1-\varepsilon) \le \varepsilon^2 \ln 4$ for all $\varepsilon \in [0,1]$.

Loss function	$H(\eta)$	$\delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y)$
Least squares	$(2\eta-1)^2$	ε^2
Hinge loss	$\lvert 2\eta-1\rvert$	ε
Squared hinge	$(2\eta-1)^2$	ε^2
Logistic loss	$\ln 2 + \Lambda(\eta) + \Lambda(1-\eta)$	$\frac{1}{2}\big(\Lambda(1+\varepsilon) + \Lambda(1-\varepsilon)\big)$

Finally, if L is classification calibrated, we have already seen $\tilde{\delta}(\varepsilon) > 0$ for all $\varepsilon \in (0,1]$, and hence $\tilde{\delta}^{**}(\varepsilon) > 0$, $\varepsilon \in (0,1]$, by Lemma A.6.20. Since we have also proved $\delta^{**} = \tilde{\delta}^{**}$, we finally find $\delta^{**}(\varepsilon) > 0$, $\varepsilon \in (0,1]$. □

For classification calibrated margin-based losses L, the preceding theorem shows that using δ^{**} in Theorem 3.22 always gives non-trivial inequalities between the excess L-risk and the excess classification risk. Furthermore, Theorem 3.34 shows that in order to establish such inequalities it suffices to compute the function $H(\,\cdot\,)$ defined by (3.37), and as we will see later in Theorem 3.36, this computation is rather simple if L is convex. For the margin-based losses considered in the examples of Section 2.3, the functions H and $\delta^{**}_{\max}(\,\cdot\,, \mathcal{Q}_Y)$ are summarized in Table 3.1. Establishing the resulting inequalities is left as an exercise (see Exercise 3.9). However, note that for some losses these inequalities can be improved if the considered P satisfies an additional assumption, as the following remark shows (see also Theorem 8.29).

Remark 3.35. It is important to note that (3.9) can be used to describe the classification scenario by a detection loss. Indeed, if for a given distribution P on $X \times Y$ with $\eta(x) := \mathrm{P}(y=1|x)$, $x \in X$, we define

$$L_{\mathrm{P}}(x,t) := |2\eta(x)-1| \cdot \mathbf{1}_{(-\infty,0]}\big((2\eta(x)-1)\operatorname{sign} t\big)\,, \qquad x \in X\,, t \in \mathbb{R},$$

then $L_{\mathrm{P}} : X \times \mathbb{R} \to [0,\infty)$ is obviously a detection loss with respect to $A := \{(x,t) \in X \times \mathbb{R} : (2\eta(x)-1)\operatorname{sign} t \le 0\}$ and $h(x) = |2\eta(x)-1|$, $x \in X$. Furthermore, (3.9) states that

$$\mathcal{C}_{L_{\mathrm{class}},\eta(x)}(t) - \mathcal{C}^*_{L_{\mathrm{class}},\eta(x)} \;=\; \mathcal{C}_{L_{\mathrm{P}},x}(t) - \mathcal{C}^*_{L_{\mathrm{P}},x}$$

for all $x \in X$, $t \in \mathbb{R}$, i.e., for the distribution P, both losses describe the same learning goal. Now, condition (3.34) becomes

$$\mathrm{P}_X\big(\{x \in X : 0 < |2\eta(x) - 1| < s\}\big) \;\le\; (c\,s)^\beta\,, \qquad s > 0, \tag{3.40}$$

which, in a slightly stronger form, will be very important condition on P when establishing fast learning rates for SVMs in Section 8.3. For now, however, we would only like to mention that, assuming (3.40), we can immediately improve the inequalities that we would obtain by combining Theorem 3.34 with Theorem 3.22 for most of the margin-based losses considered in the examples. For more details, we refer to Remark 3.30 and Exercise 3.9. ◁

Up to now, we only know that the few examples listed in Table 3.1 are classification calibrated. The following theorem gives a powerful yet easy tool to check whether a *convex* margin-based loss is classification calibrated or not.

Theorem 3.36 (Test for classification calibration). *Let L be a convex, margin-based loss represented by $\varphi : \mathbb{R} \to [0,\infty)$. Then the following statements are equivalent:*

i) L is classification calibrated.
ii) φ is differentiable at 0 and $\varphi'(0) < 0$.

Furthermore, if L is classification calibrated, then the Fenchel-Legendre biconjugate of the uniform calibration function $\delta_{\max}(\,\cdot\,, \mathcal{Q}_Y)$ satisfies

$$\delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y) = \varphi(0) - \mathcal{C}^*_{L,\frac{\varepsilon+1}{2}}\,, \qquad \varepsilon \in [0,1]. \tag{3.41}$$

Proof. ii) $\Rightarrow$ i). Since φ is differentiable at 0, the map $t \to \mathcal{C}_{L,\eta}(t)$ is differentiable at 0 and its derivative is $\mathcal{C}'_{L,\eta}(0) = (2\eta-1)\varphi'(0)$. Consequently, we have $\mathcal{C}'_{L,\eta}(0) < 0$ for $\eta \in (1/2, 1]$. Now recall that the convexity of $\mathcal{C}_{L,\eta}(\,\cdot\,)$ implies that its derivative is almost everywhere defined and increasing by Theorem A.6.6 and Proposition A.6.12. Therefore, $\mathcal{C}_{L,\eta}(\,\cdot\,)$ is decreasing on $(-\infty, 0]$ and for $\eta \in (1/2, 1]$ we thus have

$$H(\eta) = \inf_{\substack{t\in\mathbb{R}:\\ (2\eta-1)t\le 0}} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} = \inf_{t\le 0} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} = \mathcal{C}_{L,\eta}(0) - \mathcal{C}^*_{L,\eta}\,. \tag{3.42}$$

Furthermore, $\mathcal{C}'_{L,\eta}(0) < 0$ shows that $\mathcal{C}_{L,\eta}(\,\cdot\,)$ does not have a minimum at 0 and thus we find $H(\eta) > 0$ for all $\eta \in (1/2, 1]$. Lemma 3.33 then gives the classification calibration.

i) $\Rightarrow$ ii). Recall the basic facts on subdifferentials listed in Section A.6.2. Let us begin with assuming that φ is not differentiable at 0. Then there exist $w_1, w_2 \in \partial\varphi(0)$ with $w_1 < w_2$ and $w_1 \neq -w_2$. Let us fix an η with

$$\frac{1}{2} < \eta < \frac{1}{2} + \frac{w_2 - w_1}{2|w_1 + w_2|}\,.$$

Obviously, this choice implies $\frac{1}{2}(w_2 - w_1) > |w_1 + w_2|(\eta - \frac{1}{2})$, and by the definition of the subdifferential, we further have $\varphi(t) \ge w_i t + \varphi(0)$ for $t \in \mathbb{R}$ and $i = 1, 2$. For $t > 0$, we consequently find

$$\begin{aligned}\mathcal{C}_{L,\eta}(t) = \eta\varphi(t) + (1-\eta)\varphi(-t) &\geq \eta\big(w_2 t + \varphi(0)\big) + (1-\eta)\big(-w_1 t + \varphi(0)\big) \\ &= \left(\frac{1}{2}(w_2 - w_1) + (w_1 + w_2)\Big(\eta - \frac{1}{2}\Big)\right) t + \varphi(0) \\ &> \left(\Big(|w_1 + w_2| + (w_1 + w_2)\Big)\Big(\eta - \frac{1}{2}\Big)\right) t + \mathcal{C}_{L,\eta}(0) \\ &\geq \mathcal{C}_{L,\eta}(0)\,. \end{aligned} \tag{3.43}$$

Furthermore, since L is classification calibrated, we have $H(\eta) > 0$, and thus we find $\inf_{t>0} \mathcal{C}_{L,\eta}(t) = \mathcal{C}^*_{L,\eta}$. Together with (3.43), this shows $\mathcal{C}^*_{L,\eta} \geq \mathcal{C}_{L,\eta}(0)$. However, the latter yields $H(\eta) \leq 0$ by (3.42), and thus φ must be differentiable at 0. Let us now assume that $\varphi'(0) \geq 0$. We then obtain

$$\mathcal{C}_{L,1}(t) = \varphi(t) \geq \varphi'(0)t + \varphi(0) \geq \mathcal{C}_{L,1}(0)$$

for all $t > 0$. Again this contradicts the classification calibration of L.

In order to show (3.41), we first observe $\mathcal{C}'_{L,1/2}(0) = \frac{1}{2}\varphi'(0) - \frac{1}{2}\varphi'(0) = 0$. This immediately gives $\mathcal{C}_{L,1/2}(0) = \mathcal{C}^*_{L,1/2}$, and consequently we have

$$H(\eta) = \varphi(0) - \mathcal{C}^*_{L,\eta}\,, \qquad \eta \in [1/2, 1], \tag{3.44}$$

by (3.42) and $\mathcal{C}_{L,\eta}(0) = \varphi(0)$. Now recall that $\eta \to \mathcal{C}^*_{L,\eta}$ is defined by an infimum taken over affine linear functions, and hence it is a concave function. Consequently, H is convex on $[1/2, 1]$ and therefore (3.44) together with Theorem 3.34 and the continuity of L shows (3.41). □

3.5 Surrogates for Weighted Binary Classification

In this section, we investigate surrogate loss functions for the weighted binary classification scenario introduced in Example 2.5. To this end, recall that this scenario is characterized by the loss function

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

where $\alpha \in (0,1)$ was a fixed weighting parameter and $Y := \{-1, 1\}$. Adopting the notations around (3.36), we begin by computing $\delta_{\max}(\varepsilon, \eta)$.

Lemma 3.37 (Calibration function). *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss. Then, for all $\alpha \in (0,1)$, $\eta \in [0,1]$, and $\varepsilon \in (0,\infty]$, we have*

$$\delta_{\max, L_{\alpha\text{-class}}, L}(\varepsilon, \eta) = \begin{cases} \infty & \text{if } \varepsilon > |\eta - \alpha| \\ \inf_{t \in \mathbb{R}:(\eta-\alpha)\operatorname{sign} t \leq 0} \big(\mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta}\big) & \text{if } \varepsilon \leq |\eta - \alpha|. \end{cases}$$

Proof. For $t \in \mathbb{R}$, we have $\mathcal{C}_{L_{\alpha\text{-class}},\eta}(t) = (1-\alpha)\eta \mathbf{1}_{(-\infty,0)}(t) + \alpha(1-\eta)\mathbf{1}_{[0,\infty)}(t)$ and $\mathcal{C}^*_{L_{\alpha\text{-class}},\eta} = \min\{(1-\alpha)\eta, \alpha(1-\eta)\}$. From this we easily deduce

$$\mathcal{C}_{L_{\alpha\text{-class}},\eta}(t) - \mathcal{C}^*_{L_{\alpha\text{-class}},\eta} = |\eta - a| \cdot \mathbf{1}_{(-\infty,0]}\big((\eta-\alpha)\operatorname{sign} t\big)\,.$$

Now the assertion follows as in the proof of Lemma 3.32. □

In the following, we investigate how margin-based losses must be modified to make them $L_{\alpha\text{-class}}$-calibrated. To this end, let L be a margin-based loss represented by some $\varphi : \mathbb{R} \to [0,\infty)$. For $\alpha \in (0,1)$, we define the **α-weighted version** L_α of L by

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

Our next goal is to translate the results from the previous section for the unweighted classification scenario into results for the weighted case. To this end, we will frequently use the quantities

$$w_\alpha(\eta) := (1-\alpha)\eta + \alpha(1-\eta) \tag{3.45}$$

and

$$\vartheta_\alpha(\eta) := \frac{(1-\alpha)\eta}{(1-\alpha)\eta + \alpha(1-\eta)}\,, \tag{3.46}$$

which are defined for $\eta \in [0,1]$. Moreover, we need the following lemma, which describes the relation between the inner risks of L_α and L.

Lemma 3.38 (Weighted inner risks). *Let L be a margin-based loss. Then for $\alpha \in (0,1)$ and $\eta \in [0,1]$ the following statements are true:*

i) $\mathcal{C}_{L_\alpha,\eta}(t) = w_\alpha(\eta)\,\mathcal{C}_{L,\vartheta_\alpha(\eta)}(t)$ *for all* $t \in \mathbb{R}$, *and* $\mathcal{C}^*_{L_\alpha,\eta} = w_\alpha(\eta)\,\mathcal{C}^*_{L,\vartheta_\alpha(\eta)}$.
ii) $\min\{\alpha, 1-\alpha\} \le w_\alpha(\eta) \le \max\{\alpha, 1-\alpha\}$.
iii) If L is classification calibrated and $\eta \ne \alpha$, then $\mathcal{C}_{L_\alpha,\eta}(0) > \mathcal{C}^*_{L_\alpha,\eta}$.

Proof. i). A straightforward calculation shows $1-\vartheta_\alpha(\eta) = \frac{\alpha(1-\eta)}{(1-\alpha)\eta+\alpha(1-\eta)}$, and hence we obtain

$$\begin{aligned} \mathcal{C}_{L_\alpha,\eta}(t) &= (1-\alpha)\eta\varphi(t) + \alpha(1-\eta)\varphi(-t) \\ &= \big((1-\alpha)\eta + \alpha(1-\eta)\big)\big(\vartheta_\alpha(\eta)\varphi(t) + (1-\vartheta_\alpha(\eta))\varphi(-t)\big) \\ &= w_\alpha(\eta)\,\mathcal{C}_{L,\vartheta_\alpha(\eta)}(t)\,. \end{aligned}$$

ii). This follows from $w_\alpha(\eta) = (1-2\alpha)\eta + \alpha$.

iii). We have $\eta \ne \alpha$ if and only if $\vartheta_\alpha(\eta) \ne 1/2$. Furthermore, Lemma 3.33 showed $H(\eta) > 0$ for $\eta \ne 1/2$, where H is defined by (3.37), and hence we have $\mathcal{C}_{L,\eta}(0) > \mathcal{C}^*_{L,\eta}$ for $\eta \ne 1/2$. Therefore, the assertion follows from

$$\mathcal{C}_{L_\alpha,\eta}(0) = w_\alpha(\eta)\,\mathcal{C}_{L,\vartheta_\alpha(\eta)}(0) > w_\alpha(\eta)\,\mathcal{C}^*_{L,\vartheta_\alpha(\eta)} = \mathcal{C}^*_{L_\alpha,\eta}\,.$$

□

With the help of the preceding lemma, we can now characterize when α-weighted versions of margin-based loss functions are $L_{\alpha\text{-class}}$-calibrated.

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*

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

Then the following statements are equivalent:

i) L_α is uniformly $L_{\alpha\text{-class}}$-calibrated with respect to $\mathcal{Q}_Y$.
ii) L_α is $L_{\alpha\text{-class}}$-calibrated with respect to $\mathcal{Q}_Y$.
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

$$H_\alpha(\eta) = w_\alpha(\eta) H\big(\vartheta_\alpha(\eta)\big)\,. \tag{3.48}$$

Proof. ii) $\Leftrightarrow$ iii). An easy calculation shows $2\vartheta_\alpha(\eta) - 1 = \frac{\eta-\alpha}{(1-\alpha)\eta+\alpha(1-\eta)}$, and hence we find $\operatorname{sign}(\eta-\alpha) = \operatorname{sign}(2\vartheta_\alpha(\eta)-1)$. For $\varepsilon \le |\eta-\alpha|$, this gives

$$\begin{aligned}
\delta_{\max,L_{\alpha\text{-class}},L_\alpha}(\varepsilon,\eta) &= \inf_{\substack{t\in\mathbb{R}\\ (\eta-\alpha)\operatorname{sign} t\le 0}} \mathcal{C}_{L_\alpha,\eta}(t) - \mathcal{C}^*_{L_\alpha,\eta} \\
&= w_\alpha(\eta) \inf_{\substack{t\in\mathbb{R}\\ (2\vartheta_\alpha(\eta)-1)\operatorname{sign} t\le 0}} \mathcal{C}_{L,\vartheta_\alpha(\eta)}(t) - \mathcal{C}^*_{L,\vartheta_\alpha(\eta)} \\
&= w_\alpha(\eta)\, \delta_{\max,L_{\text{class}},L}(\varepsilon,\vartheta_\alpha(\eta))\,.
\end{aligned} \tag{3.49}$$

Since $w_\alpha(\eta) > 0$ and $\vartheta_\alpha([0,1]) = [0,1]$, we then obtain the equivalence. The proof of (3.48) is analogous to (3.49).

i) $\Rightarrow$ ii). Trivial.

iii) $\Rightarrow$ i). Recall that L is uniformly classification calibrated by Theorem 3.34. Then the implication follows from using $w_\alpha(\eta) \ge \min\{\alpha, 1-\alpha\}$ in (3.49).

ii) $\Rightarrow$ iv). Part *iii)* of Lemma 3.38 together with Lemma 3.37 implies $H_\alpha(\eta) > 0$ for all $\eta \neq \alpha$.

iv) $\Rightarrow$ ii). By Lemma 3.37, we have $\delta_{\max,\alpha}(\varepsilon,\eta) \ge H_\alpha(\eta) > 0$ for $\eta \neq \alpha$ and $0 < \varepsilon \le |\eta-\alpha|$. This gives the assertion. □

With the help of the results above we can now establish our main theorem of this section, which describes an easy way to establish inequalities for $L_{\alpha\text{-class}}$-calibrated loss functions.

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

$$\delta_\alpha(\varepsilon) := \inf_{\substack{\eta\in[0,1]\\ |\eta-\alpha|\ge\varepsilon}} H_\alpha(\eta)\,, \qquad \varepsilon \in [0,\alpha_{\max}],$$

Table 3.2. The functions H, H_α, and δ_α^{**} for some common margin-based losses. The values for δ_α^{**} are only for α with $0 < \alpha \le 1/2$. Note that, for the hinge loss, the function δ_α^{**} is actually independent of α. Furthermore, the formulas for the logistic loss for classification do not fit into the table but can be easily computed.

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	$\lvert 2\eta-1\rvert$	$\lvert\eta-\alpha\rvert$	ε
Squared hinge	$(2\eta-1)^2$	$\frac{(\eta-\alpha)^2}{\alpha+\eta-2\alpha\eta}$	$\frac{\varepsilon^2}{2\alpha(1-\alpha)+\varepsilon(1-2\alpha)}$

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

$$\delta_\alpha^{**}(\varepsilon) \le \delta^{**}_{\max, L_{\alpha\text{-class}}, L}(\varepsilon, \mathcal{Q}_Y)\,,$$

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

Proof. Let $\varepsilon \in [0, \alpha_{\max}]$. Then Lemma 3.37 together with $\inf \emptyset = \infty$ yields

$$\inf_{Q\in\mathcal{Q}_Y} \delta_{\max}(\varepsilon, Q) = \inf_{\substack{\eta\in[0,1]\\ |\eta-\alpha|\ge\varepsilon}} \inf_{\substack{t\in\mathbb{R}\\ (\eta-\alpha)\operatorname{sign} t\le 0}} \mathcal{C}_{L_\alpha,\eta}(t) - \mathcal{C}^*_{L_\alpha,\eta} \ge \inf_{\substack{\eta\in[0,1]\\ |\eta-\alpha|\ge\varepsilon}} H_\alpha(\eta)\,.$$

□

Obviously, we can use the identity $H_\alpha(\eta) = w_\alpha(\eta) H\big(\vartheta_\alpha(\eta)\big)$ in order to compute the function $\delta_\alpha(\varepsilon)$ of the preceding theorem. Doing so, we see that δ_α is a continuous function that is strictly positive on $(0, \alpha_{\max}]$ if L is classification calibrated. Consequently, Theorem 3.40 together with Theorem 3.22 yields non-trivial inequalities. Furthermore, for some important loss functions, we already know $H(\eta)$, $\eta \in [0,1]$, and hence the computation of $\delta_\alpha^{**}(\varepsilon)$ is straightforward. The corresponding results are summarized in Table 3.2.

Up to now, we have only investigated the $L_{\alpha\text{-class}}$-calibration of α-weighted versions of classification calibrated loss functions. We finally show that other weighted versions are *not* $L_{\alpha\text{-class}}$-calibrated.

Theorem 3.41 (Using the correct weights). *Let $\alpha, \beta \in (0,1)$, L be a margin-based, classification calibrated loss, and L_β be its β-weighted version. Then L_β is $L_{\alpha\text{-class}}$-calibrated if and only if $\beta = \alpha$.*

Proof. We already know that L_α is $L_{\alpha\text{-class}}$-calibrated, and hence we assume $\alpha \ne \beta$. Without loss of generality, we only consider the case $\beta > \alpha$. For a fixed $\eta \in (\alpha, \beta)$, an easy computation then shows that $\vartheta_\beta(\eta)$ defined in (3.46) satisfies $\vartheta_\beta(\eta) < 1/2 < \vartheta_\alpha(\eta)$, and hence for $\varepsilon > 0$ with $\varepsilon \le |\eta - \alpha|$ we obtain

$$\begin{aligned}\delta_{\max, L_{\alpha\text{-class}}, L_\beta}(\varepsilon, \eta) &= \inf_{(n-\alpha)\operatorname{sign} t\le 0} \mathcal{C}_{L_\beta,\eta}(t) - \mathcal{C}^*_{L_\beta,\eta} \\ &= w_\beta(\eta) \inf_{t<0} \mathcal{C}_{L,\vartheta_\beta(\eta)}(t) - \mathcal{C}^*_{L,\vartheta_\beta(\eta)}\,. \qquad (3.50)\end{aligned}$$

The classification calibration of L implies $\inf_{t\geq 0} \mathcal{C}_{L,\vartheta_\beta(\eta)}(t) - \mathcal{C}^*_{L,\vartheta_\beta(\eta)} > 0$, and since $\inf_{t\in\mathbb{R}} \mathcal{C}_{L,\vartheta_\beta(\eta)}(t) - \mathcal{C}^*_{L,\vartheta_\beta(\eta)} = 0$, we find $\inf_{t<0} \mathcal{C}_{L,\vartheta_\beta(\eta)}(t) - \mathcal{C}^*_{L,\vartheta_\beta(\eta)} = 0$. Together with (3.50), this shows that L_β is not $L_{\alpha\text{-class}}$-calibrated. □

The preceding theorem in particular shows that an α-weighted version of a classification calibrated margin-based loss function is classification calibrated if and only if $\alpha = 1/2$. In other words, using a weighted margin-based loss for an unweighted classification problem may lead to methodical errors.

3.6 Template Loss Functions

Sometimes an unsupervised loss function explicitly depends on the data-generating distribution. For example, if we have a distribution P on $X \times \mathbb{R}$ with $|\mathrm{P}|_1 < \infty$ and we wish to estimate the conditional mean function $x \mapsto \mathbb{E}_\mathrm{P}(Y|x)$, we could describe this learning goal by the loss function

$$L(x,t) := \big|\mathbb{E}_\mathrm{P}(Y|x) - t\big|\,, \qquad\qquad x \in X,\, t \in \mathbb{R}.$$

Now note that when we change the distribution we have to change the loss function, though the learning goal remains the same. In view of our analysis on surrogate losses, this fact is at least annoying. The goal of this section is to resolve this issue by introducing a new type of "loss function" that may depend on distributions Q. Let us begin with a precise definition.

Definition 3.42. *Let $\mathcal{Q}$ be a set of distributions on a closed subset $Y \subset \mathbb{R}$. Then we call a function $L : \mathcal{Q} \times \mathbb{R} \to [0,\infty)$ a* ***template loss*** *if, for all complete measurable spaces X and all distributions* P *of type $\mathcal{Q}$, the* P***-instance*** *L_P of L defined by*

$$\begin{aligned} L_\mathrm{P} : X \times \mathbb{R} &\to [0,\infty) \\ (x,t) &\mapsto L\big(\mathrm{P}(\,\cdot\,|x), t\big) \end{aligned} \tag{3.51}$$

is measurable.

Note that the key condition of this definition is the *measurability*, which enables us to interpret P-instances as unsupervised losses. In particular, we can define the risk of a template loss L by the risk of its P-instance, i.e., by

$$\mathcal{R}_{L,\mathrm{P}}(f) := \mathcal{R}_{L_\mathrm{P},\mathrm{P}}(f) = \int_X L\big(\mathrm{P}(\,\cdot\,|x), f(x)\big)\, d\mathrm{P}_X(x)\,,$$

where $f : X \to \mathbb{R}$ is measurable. This motivates us to define the inner risks of a template loss $L : \mathcal{Q} \times \mathbb{R} \to [0,\infty)$ analogously to the inner risks of unsupervised losses, i.e., we write

$$\begin{aligned} \mathcal{C}_{L,\mathrm{Q}}(t) &:= L(\mathrm{Q},t), \\ \mathcal{C}^*_{L,\mathrm{Q}} &:= \inf_{t'\in\mathbb{R}} L(\mathrm{Q},t') \end{aligned}$$

for $\mathrm{Q} \in \mathcal{Q}$ and $t \in \mathbb{R}$. Note that the right-hand sides of these definitions have the form we used for unsupervised losses in the sense that no integrals occur while the left-hand sides have the form we obtained for supervised losses in the sense that the inner risks are independent of x. Having defined the inner risks, we write, as usual,

$$\mathcal{M}_{L,\mathrm{Q}}(\varepsilon) := \left\{ t \in \mathbb{R} : \mathcal{C}_{L,\mathrm{Q}}(t) < \mathcal{C}^*_{L,\mathrm{Q}} + \varepsilon \right\}, \qquad \mathrm{Q} \in \mathcal{Q},\ \varepsilon \in [0,\infty],$$

for the corresponding sets of approximate minimizers. Moreover, given a *supervised* surrogate loss $L_{\mathrm{sur}} : Y \times \mathbb{R} \to [0,\infty)$, we define the calibration function $\delta_{\max}(\,\cdot\,, \mathrm{Q}) : [0,\infty] \to [0,\infty]$ of (L, L_{sur}) by

$$\delta_{\max,L,L_{\mathrm{sur}}}(\varepsilon, \mathrm{Q}) := \inf_{\substack{t \in \mathbb{R} \\ t \notin \mathcal{M}_{L,\mathrm{Q}}(\varepsilon)}} \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q}}(t) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q}}, \qquad \varepsilon \in [0,\infty],$$

if $\mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q}} < \infty$ and by $\delta_{\max,L,L_{\mathrm{sur}}}(\varepsilon, \mathrm{Q}) := \infty$ otherwise. Since in the proof of Lemma 3.14 we did not use that the inner risks are defined by integrals, it is then not hard to see that this lemma also holds for the calibration function above. Consequently, we say that L_{sur} is **L-calibrated** with respect to $\mathcal{Q}$ if

$$\delta_{\max,L,L_{\mathrm{sur}}}(\varepsilon, \mathrm{Q}) > 0$$

for all $\varepsilon > 0$ and $\mathrm{Q} \in \mathcal{Q}$. Analogously, we say that L_{sur} is **uniformly L-calibrated** with respect to $\mathcal{Q}$ if

$$\delta_{\max,L,L_{\mathrm{sur}}}(\varepsilon, \mathcal{Q}) := \inf_{\mathrm{Q} \in \mathcal{Q}} \delta_{\max,L,L_{\mathrm{sur}}}(\varepsilon, \mathrm{Q}) > 0$$

for all $\varepsilon > 0$. If we now consider a P-instance L_{P} of L, we immediately obtain

$$\delta_{\max,L_{\mathrm{P}},L_{\mathrm{sur}}}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) = \delta_{\max,L,L_{\mathrm{sur}}}(\varepsilon, \mathrm{P}(\,\cdot\,|x)) \tag{3.52}$$

for all $\varepsilon \in [0,\infty]$ and $x \in X$, where $\delta_{\max,L_{\mathrm{P}},L_{\mathrm{sur}}}(\,\cdot\,,\,\cdot\,,\,\cdot\,)$ denotes the calibration function of $(L_{\mathrm{P}}, L_{\mathrm{sur}})$. In other words, L-calibration of L_{sur} can be investigated analogously to *supervised* losses, i.e., in terms of $\mathcal{Q}$ and independent of x, while the corresponding results can be used to determine the relation between the excess L_{sur}-risk and the excess risk of the *unsupervised* loss L_{P}. In the following sections, we will extensively make use of template losses, mainly because of this technical merit.

3.7 Surrogate Losses for Regression Problems

In regression, the goal is to predict a real-valued output y given an input x. The discrepancy between the prediction $f(x)$ and the observation y is often measured by the least squares loss, but we have already seen in Section 2.4 that there are various alternatives. In this section, we investigate the relation

of these alternatives to the least squares loss. These considerations will be important for Chapters 9 and 10 on regression and robustness, respectively.

Let us begin by introducing some notation. To this end let, $\mathcal{Q}$ be a set of distributions on $\mathbb{R}$ and $L : \mathbb{R} \times \mathbb{R} \to [0, \infty]$ be a supervised loss. Since in our general results on calibration the assumption $\mathcal{C}^*_{L,\mathrm{Q}} < \infty$ was crucial, we define

$$\mathcal{Q}(L) := \left\{ \mathrm{Q} \in \mathcal{Q} : \mathcal{C}^*_{L,\mathrm{Q}} < \infty \right\}.$$

Recall that for distance-based losses we have investigated the condition $\mathcal{C}^*_{L,\mathrm{Q}} < \infty$ in Lemma 2.36. In the following, $\mathcal{Q}_{\mathbb{R}}$ denotes the set of distributions on $\mathbb{R}$, and more generally, $\mathcal{Q}_I$ denotes the set of all distributions whose support is contained in the subset $I \subset \mathbb{R}$. In addition, for $p \in (0, \infty]$, the set of distributions on $\mathbb{R}$ with p-th finite moment is denoted by

$$\mathcal{Q}^{(p)}_{\mathbb{R}} := \left\{ \mathrm{Q} : \mathrm{Q} \text{ distribution on } \mathbb{R} \text{ with } |\mathrm{Q}|_p < \infty \right\},$$

whereas the set of all distributions with bounded support is denoted by

$$\mathcal{Q}_{\text{bounded}} := \mathcal{Q}^{(\infty)}_{\mathbb{R}} = \bigcup_{M>0} \mathcal{Q}_{[-M,M]}.$$

Note that $\mathcal{Q}_I \subset \mathcal{Q}_{\text{bounded}} \subset \mathcal{Q}^{(1)}_{\mathbb{R}}$ holds for all bounded intervals I, and if L is a continuous, distance-based loss, we actually have $\mathcal{Q}_{\text{bounded}} \subset \mathcal{Q}^{(1)}_{\mathbb{R}}(L)$.

Now let Q be a distribution on $\mathbb{R}$ such that $|\mathrm{Q}|_1 < \infty$. Then the **mean** of Q is denoted by

$$\mathbb{E}\mathrm{Q} := \int_{\mathbb{R}} y \, d\mathrm{Q}(y).$$

We call Q **symmetric** around some $c \in \mathbb{R}$ if $\mathrm{Q}(c + A) = \mathrm{Q}(c - A)$ for all measurable $A \subset [0, \infty)$. Furthermore, we say that Q is symmetric if it is symmetric around some $c \in \mathbb{R}$. Obviously, Q is symmetric around c if and only if its **centered version** $\mathrm{Q}^{(c)}$ defined by $\mathrm{Q}^{(c)}(A) := \mathrm{Q}(c + A)$, $A \subset \mathbb{R}$ measurable, is centered around 0. In the following, the set of all symmetric distributions with p-finite moment is denoted by $\mathcal{Q}^{(p)}_{\mathbb{R},\text{sym}}$. Finally, the sets $\mathcal{Q}_{I,\text{sym}}$, for $I \subset \mathbb{R}$, and $\mathcal{Q}_{\text{bounded,sym}}$ are defined in the obvious way.

Let us now assume that Q is symmetric around c. For a measurable function $h : \mathbb{R} \to \mathbb{R}$, we then have

$$\begin{aligned} \int_{\mathbb{R}} h(y - c) d\mathrm{Q}(y) \;&=\; \int_{\mathbb{R}} h(y) d\mathrm{Q}^{(c)}(y) = \int_{\mathbb{R}} h(-y) d\mathrm{Q}^{(c)}(y) \\ &= \int_{\mathbb{R}} h(c - y) d\mathrm{Q}(y) \qquad (3.53) \end{aligned}$$

whenever one (and then all) of the integrals exists. In particular, for $h(y) := y + c$, $y \in \mathbb{R}$, and Q satisfying $|\mathrm{Q}|_1 < \infty$, this equation yields

$$\mathbb{E}\mathrm{Q} = \int_{\mathbb{R}} y d\mathrm{Q}(y) = \int_{\mathbb{R}} h(y - c) d\mathrm{Q}(y) = c + \int_{\mathbb{R}} y d\mathrm{Q}^{(c)}(y) = c,$$

i.e., the center c is unique and equals the mean $\mathbb{E}\mathrm{Q}$.

Let us get back to our main goal, which is identifying L_{LS}-calibrated losses, where L_{LS} denotes the least squares loss. To this end, recall that for $\mathrm{Q} \in \mathcal{Q}_{\mathbb{R}}(L_{\mathrm{LS}}) = \mathcal{Q}_{\mathbb{R}}^{(2)}$ we have already seen in Example 2.6 that

$$\mathcal{M}_{L_{\mathrm{LS}},\mathrm{Q}}(0^+) = \{\mathbb{E}\mathrm{Q}\}\,.$$

Consequently, if L is a supervised, L_{LS}-calibrated loss function, we must have $\mathcal{M}_{L,\mathrm{Q}}(0^+) \subset \{\mathbb{E}\mathrm{Q}\}$ for all $\mathrm{Q} \in \mathcal{Q}_{\mathbb{R}}^{(2)}(L)$. This observation motivates the following two propositions in which we investigate the sets $\mathcal{M}_{L,\mathrm{Q}}(0^+)$ for distance-based losses.

Proposition 3.43 (Exact minimizers for distance-based losses I). *Let L be a distance-based loss whose representing function $\psi : \mathbb{R} \to [0,\infty)$ satisfies $\lim_{r\to\pm\infty} \psi(r) = \infty$. Moreover, let $\mathrm{Q} \in \mathcal{Q}_{\mathbb{R}}$ be a distribution with $\mathcal{C}_{L,\mathrm{Q}}(t) < \infty$ for all $t \in \mathbb{R}$. Then the following statements are true:*

i) If ψ is convex, then $t \mapsto \mathcal{C}_{L,\mathrm{Q}}(t)$ is convex and continuous. Moreover, we have $\lim_{t\to\pm\infty} \mathcal{C}_{L,\mathrm{Q}}(t) = \infty$ and $\mathcal{M}_{L,\mathrm{Q}}(0^+) \neq \emptyset$.

ii) If ψ is strictly convex, then $t \mapsto \mathcal{C}_{L,\mathrm{Q}}(t)$ is strictly convex and $\mathcal{M}_{L,\mathrm{Q}}(0^+)$ contains exactly one element.

Proof. Our first goal is to show that $\lim_{t\to\pm\infty} \mathcal{C}_{L,\mathrm{Q}}(t) = \infty$. To this end, we fix a $B > 0$ and let $(t_n) \subset \mathbb{R}$ be a sequence with $t_n \to -\infty$. Since $\lim_{r\to\pm\infty} \psi(r) = \infty$, there then exists an $r_0 > 0$ such that $\psi(r) \geq 2B$ for all $r \in \mathbb{R}$ with $|r| \geq r_0$. Since $\mathrm{Q}(\mathbb{R}) = 1$, there exists also an $M > 0$ with $\mathrm{Q}([-M,M]) \geq 1/2$. Finally, there exists an $n_0 \geq 1$ with $t_n \leq -M - r_0$ for all $n \geq n_0$. For $y \in [-M,M]$, this yields $y - t_n \geq r_0$, and hence we find $\psi(y - t_n) \geq 2B$ for all $n \geq n_0$. From this we easily conclude

$$\mathcal{C}_{L,\mathrm{Q}}(t_n) \geq \int_{[-M,M]} \psi(y - t_n)\, d\mathrm{Q}(y) \geq 2B\, \mathrm{Q}\big([-M,M]\big) = B\,,$$

i.e., we have shown $\mathcal{C}_{L,\mathrm{Q}}(t_n) \to \infty$. Analogously we can show $\lim_{t\to\infty} \mathcal{C}_{L,\mathrm{Q}}(t) = \infty$, and consequently we have $\lim_{t\to\pm\infty} \mathcal{C}_{L,\mathrm{Q}}(t) = \infty$. This shows that

$$\big\{t \in \mathbb{R} : \mathcal{C}_{L,\mathrm{Q}}(t) \leq \mathcal{C}_{L,\mathrm{Q}}(0)\big\}$$

is a non-empty and bounded subset of $\mathbb{R}$. Furthermore, the convexity of ψ implies that $t \mapsto \mathcal{C}_{L,\mathrm{Q}}(t)$ is convex and hence this map is continuous by Lemma A.6.2. Now the assertions follow from Theorem A.6.9. □

Note that for distributions $\mathrm{Q} \in \mathcal{Q}_{\mathrm{bounded}}$ we automatically have $\mathcal{C}_{L,\mathrm{Q}}(t) < \infty$ for all $t \in \mathbb{R}$ and all distance-based losses L. Furthermore, if L is of some growth type $p \in (0,\infty)$, then Lemma 2.36 shows $\mathcal{C}_{L,\mathrm{Q}}(t) < \infty$ for all $t \in \mathbb{R}$ and all distributions Q having finite p-th moment. Consequently, the preceding proposition gives $\mathcal{M}_{L,\mathrm{Q}}(0^+) \neq \emptyset$ in both cases.

The following proposition compares $\mathcal{M}_{L,\mathrm{Q}}(0^+)$ with the mean $\mathbb{E}\mathrm{Q}$.

Proposition 3.44 (Exact minimizers for distance-based losses II). *Let L be a distance-based loss whose representing function ψ is locally Lipschitz continuous, and let $M > 0$. Then the following statements are true:*

i) If $\mathbb{E}\mathrm{Q} \in \mathcal{M}_{L,\mathrm{Q}}(0^+)$ for all $\mathrm{Q} \in \mathcal{Q}_{[-M,M],\mathrm{sym}}$, then L is symmetric.
ii) If $\mathbb{E}\mathrm{Q} \in \mathcal{M}_{L,\mathrm{Q}}(0^+)$ for all $\mathrm{Q} \in \mathcal{Q}_{\mathrm{bounded}}$, then there exists a constant $c \geq 0$ with $\psi(t) = ct^2$ for all $t \in \mathbb{R}$.

Proof. Recall that the fundamental theorem of calculus for Lebesgue integrals (see Theorem A.6.6) shows that the derivative ψ' is (Lebesgue)-almost surely defined and integrable on every bounded interval.

i). Let us fix a $y \in [-M, M]$ such that ψ is differentiable at y and $-y$. We define $\mathrm{Q} := \frac{1}{2}\delta_{\{-y\}} + \frac{1}{2}\delta_{\{y\}}$. Then we have $\mathrm{Q} \in \mathcal{Q}_{\mathbb{R},\mathrm{sym}}$ with $\mathbb{E}\mathrm{Q} = 0$, and $\mathcal{C}_{L,\mathrm{Q}}(t) = \frac{1}{2}\psi(-y-t)+\frac{1}{2}\psi(y-t)$. Consequently, the derivative of $\mathcal{C}_{L,\mathrm{Q}}(\,\cdot\,)$ exists at 0 and can be computed by $\mathcal{C}'_{L,\mathrm{Q}}(0) = -\frac{1}{2}\psi'(-y) - \frac{1}{2}\psi'(y)$. Furthermore, our assumption shows that $\mathcal{C}_{L,\mathrm{Q}}(\,\cdot\,)$ has a minimum at 0, and hence we have $0 = \mathcal{C}'_{L,\mathrm{Q}}(0)$, i.e., $\psi'(-y) = -\psi'(y)$. According to our preliminary remark, the latter relation holds for almost all y, and hence Theorem A.6.6 shows that, for all $y_0 \in \mathbb{R}$, we have

$$\begin{aligned}\psi(y_0) \;=\; \psi(0) + \int_0^{y_0} \psi'(t)dt \;=\; \psi(0) - \int_0^{y_0} \psi'(-t)dt = \psi(0) - \int_{-y_0}^{0} \psi'(t)dt \\ = \psi(-y_0)\,.\end{aligned}$$

ii). Let $y \neq 0$ and $\alpha > 0$ be real numbers such that ψ is differentiable at y, $-y$, and αy. We define $\mathrm{Q} := \frac{\alpha}{1+\alpha}\delta_{\{0\}} + \frac{1}{1+\alpha}\delta_{\{(1+\alpha)y\}}$, so that we obtain $\mathbb{E}\mathrm{Q} = y$ and $\mathcal{C}_{L,\mathrm{Q}}(t) = \frac{\alpha}{1+\alpha}\psi(-t) + \frac{1}{1+\alpha}\psi(y+\alpha y-t)$ for all $t \in \mathbb{R}$. This shows that the derivative of $\mathcal{C}_{L,\mathrm{Q}}(\,\cdot\,)$ exists at y and can be computed by

$$\mathcal{C}'_{L,\mathrm{Q}}(y) = -\frac{\alpha}{1+\alpha}\psi'(-y) - \frac{1}{1+\alpha}\psi'(\alpha y) = \frac{\alpha}{1+\alpha}\psi'(y) - \frac{1}{1+\alpha}\psi'(\alpha y)\,,$$

where in the last step we used *i)*. Now, our assumption $\mathbb{E}\mathrm{Q} \in \mathcal{M}_{L,\mathrm{Q}}(0^+)$ gives $\mathcal{C}'_{L,\mathrm{Q}}(y) = 0$, and hence we find $\alpha\psi'(y) = \psi'(\alpha y)$. Obviously, the latter relation holds for almost all $\alpha > 0$, and thus we obtain

$$\psi(ty) = \psi(0) + \int_0^t \psi'(sy)y\,ds = \int_0^t s\psi'(y)y\,ds = \frac{\psi'(y)}{2y}(ty)^2$$

for all $t > 0$. From this we easily obtain the assertion for $c := \frac{\psi'(y)}{2y}$. □

Proposition 3.44 shows that there is basically *no* distance-based surrogate for the least squares loss L_{LS} if one is interested in the entire class

$$\mathcal{Q}_{\mathbb{R}}(L_{\mathrm{LS}}) = \mathcal{Q}_{\mathbb{R}}^{(2)} = \big\{\mathrm{Q} \in \mathcal{Q}_{\mathbb{R}} : |\mathrm{Q}|_2 < \infty\big\}\,.$$

Furthermore, it shows that the least squares loss is essentially the only distance-based loss function whose minimizer is the mean for all distributions

in $\mathcal{Q}_{\mathbb{R}}^{(2)}$. In other words, if we are actually interested in finding the **regression function** $x \mapsto \mathbb{E}_P(Y|x)$, and we just know $|P|_2 < \infty$, then the least squares loss is the *only* suitable distance-based loss for this task. However, if we cannot ensure the tail assumption $|P|_2 < \infty$ but know instead that the conditional distributions $P(\cdot\,|x)$ are *symmetric*, then Proposition 3.44 suggests that we may actually have alternatives to the least squares loss. In order to investigate this conjecture systematically, we first need a target loss that describes the goal of estimating the mean. To this end, let us consider the **mean distance** template loss $L_{\text{mean}} : \mathcal{Q}_{\mathbb{R}}^{(1)} \times \mathbb{R} \to [0,\infty)$, which is defined by

$$L_{\text{mean}}(Q,t) := |\mathbb{E}Q - t|\,, \qquad t \in \mathbb{R}\,, Q \in \mathcal{Q}_{\mathbb{R}}^{(1)}\,.$$

Note that this indeed defines a template loss, since given a $\mathcal{Q}_{\mathbb{R}}^{(1)}$-type distribution P on $X \times \mathbb{R}$, it is easy to see that

$$(x,t) \mapsto L_{\text{mean}}(P(\cdot\,|x),t) = \bigl|\mathbb{E}_P(Y|x) - t\bigr|$$

is measurable. Moreover, we have

$$L_{\text{mean}}^2(Q,t) = \bigl(\mathbb{E}Q - t\bigr)^2 = \mathcal{C}_{L_{\text{LS}},Q}(t) - \mathcal{C}^*_{L_{\text{LS}},Q}\,, \qquad Q \in \mathcal{Q}_{\mathbb{R}}^{(2)},\, t \in \mathbb{R},$$

and since the minimal L_{mean}-risks equal 0, we thus obtain $\mathcal{M}_{L_{\text{mean}},Q}(\sqrt{\varepsilon}) = \mathcal{M}_{L_{\text{LS}},Q}(\varepsilon)$ for all $\varepsilon > 0$. From this we immediately find

$$\delta_{\max,L_{\text{mean}},L}(\sqrt{\varepsilon},Q) = \delta_{\max,L_{\text{LS}},L}(\varepsilon,Q)\,, \qquad \varepsilon \in [0,\infty], \tag{3.54}$$

for all distance-based losses L and all $Q \in \mathcal{Q}_{\mathbb{R}}^{(2)} \cap \mathcal{Q}_{\mathbb{R}}(L)$. In other words, by considering L_{mean}-calibration, we simultaneously obtain results on L_{LS}-calibration.

We saw in Section 3.1 that the inner risks are the key quantities for computing calibration functions. The following lemma presents a way to compute the inner risks $\mathcal{C}_{L,Q}(\cdot\,)$ when both L and Q are symmetric.

Lemma 3.45 (Inner risks of symmetric losses). *Let L be a symmetric loss with representing function ψ and $Q \in \mathcal{Q}_{\mathbb{R},\text{sym}}^{(1)}(L)$. Then we have*

$$\mathcal{C}_{L,Q}(\mathbb{E}Q + t) = \mathcal{C}_{L,Q}(\mathbb{E}Q - t) = \frac{1}{2}\int_{\mathbb{R}} \psi(y - \mathbb{E}Q - t) + \psi(y - \mathbb{E}Q + t)\,dQ(y)$$

for all $t \in \mathbb{R}$. In addition, if L is convex, we have

$$\mathcal{C}_{L,Q}(\mathbb{E}Q) = \mathcal{C}^*_{L,Q}\,,$$

and if L is strictly convex, we also have $\mathcal{C}_{L,Q}(\mathbb{E}Q + t) > \mathcal{C}^*_{L,Q}$ for all $t \neq 0$.

Proof. Let us write $m := \mathbb{E}Q$. Recalling that the centered version $Q^{(m)}$ of Q is symmetric around 0, the symmetry of ψ and (3.53) then yield

$$\begin{aligned}\mathcal{C}_{L,\mathrm{Q}}(m+t) = \int_{\mathbb{R}} \psi(y-t)d\mathrm{Q}^{(m)}(y) &= \int_{\mathbb{R}} \psi(-y-t)d\mathrm{Q}^{(m)}(y) \\ &= \int_{\mathbb{R}} \psi(y+t)d\mathrm{Q}^{(m)}(y) \\ &= \mathcal{C}_{L,\mathrm{Q}}(m-t)\,.\end{aligned}$$

Since this yields $\mathcal{C}_{L,\mathrm{Q}}(m+t) = \frac{1}{2}(\mathcal{C}_{L,\mathrm{Q}}(m+t) + \mathcal{C}_{L,\mathrm{Q}}(m-t))$, we also obtain the second equation. Furthermore, if ψ is convex, we can easily conclude that

$$\mathcal{C}_{L,\mathrm{Q}}(m+t) = \frac{1}{2}\int_{\mathbb{R}} \psi(y-t)+\psi(y+t)d\mathrm{Q}^{(m)}(y) \geq \int_{\mathbb{R}} \psi(y)d\mathrm{Q}^{(m)}(y) = \mathcal{C}_{L,\mathrm{Q}}(m)$$

for all $t \in \mathbb{R}$. This shows the second assertion. The third assertion can be shown analogously. □

With the help of the preceding lemma, we can derive a simple formula for the calibration function $\delta_{\max,L_{\mathrm{mean}},L}(\varepsilon,\mathrm{Q})$ if L is convex.

Lemma 3.46 (Calibration function for symmetric losses). *Let L be a symmetric, convex loss and $\mathrm{Q} \in \mathcal{Q}^{(1)}_{\mathbb{R},\mathrm{sym}}(L)$. Then, for all $\varepsilon \geq 0$, we have*

$$\delta_{\max,L_{\mathrm{mean}},L}(\varepsilon,\mathrm{Q}) = \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+\varepsilon) - \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q})\,. \tag{3.55}$$

Consequently, $\varepsilon \mapsto \delta_{\max,L_{\mathrm{mean}},L}(\varepsilon,\mathrm{Q})$ is convex and the following statements are equivalent:

i) $\delta_{\max,L_{\mathrm{mean}},L}(\varepsilon,\mathrm{Q}) > 0$ for all $\varepsilon > 0$.
ii) $\mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+t) > \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q})$ for all $t \in \mathbb{R}$ with $t \neq 0$.

Proof. Obviously, it suffices to prove (3.55). To this end, observe that $t \mapsto \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+t)$ is a convex function on $\mathbb{R}$, and Lemma 3.45 shows that it is also symmetric in the sense of $\mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+t) = \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}-t)$ for all $t \in \mathbb{R}$. Therefore, $t \mapsto \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+t)$ is decreasing on $(\infty,0]$ and increasing on $[0,\infty)$, and hence we find

$$\delta_{\max,L_{\mathrm{mean}},L}(\varepsilon,\mathrm{Q}) = \inf_{t\notin(\varepsilon,\varepsilon)} \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+t) - \mathcal{C}^*_{L,\mathrm{Q}} = \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+\varepsilon) - \mathcal{C}^*_{L,\mathrm{Q}}\,.$$

Since we already know that $\mathcal{C}^*_{L,\mathrm{Q}} = \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q})$ by Lemma 3.45, we then obtain the assertion. □

Our next result is a technical lemma that will be used to establish *upper* bounds on $\delta_{\max,L_{\mathrm{mean}},L}(\varepsilon,\mathrm{Q})$. For its formulation, we need the set

$$\mathcal{Q}^*_{\mathbb{R},\mathrm{sym}} := \Big\{\mathrm{Q} \in \mathcal{Q}^{(1)}_{\mathbb{R},\mathrm{sym}} : \mathrm{Q}\big([\mathbb{E}\mathrm{Q}-\rho,\mathbb{E}\mathrm{Q}+\rho]\big) > 0 \text{ for all } \rho > 0\Big\}\,,$$

which contains all symmetric distributions on $\mathbb{R}$ that do not vanish around their means. Moreover, we also need the sets $\mathcal{Q}^*_{I,\mathrm{sym}} := \mathcal{Q}_I \cap \mathcal{Q}^*_{\mathbb{R},\mathrm{sym}}$, for $I \subset \mathbb{R}$, and $\mathcal{Q}^*_{\mathrm{bounded,sym}} := \mathcal{Q}_{\mathrm{bounded}} \cap \mathcal{Q}^*_{\mathbb{R},\mathrm{sym}}$. Now the result reads as follows.

Lemma 3.47 (Upper bound on excess risks). *Let L be a symmetric, continuous loss with representing function ψ. Assume that there exist a $\delta_0 \in \mathbb{R}$, $s_1, s_2 \in \mathbb{R}$ with $s_1 \neq s_2$, and an $\varepsilon_0 > 0$ such that for all $\varepsilon \in [0, \varepsilon_0]$ we have*

$$\frac{\psi(s_1+\varepsilon)+\psi(s_2+\varepsilon)}{2} - \psi\Big(\frac{s_1+s_2}{2}+\varepsilon\Big) \leq \delta_0\,. \tag{3.56}$$

Let us write $M := |\frac{s_1+s_2}{2}| + \varepsilon_0$ and $t := \frac{s_2-s_1}{2}$. Then, for all $\delta > 0$, there exists a Lebesgue absolutely continuous $\mathrm{Q} \in \mathcal{Q}^*_{[-M,M],\mathrm{sym}}$ with $\mathbb{E}\mathrm{Q} = 0$ and

$$\mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+t) - \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}) \leq \delta_0 + \delta\,.$$

Moreover, there exists a Lebesgue absolutely continuous $\mathrm{Q} \in \mathcal{Q}_{[-M,M],\mathrm{sym}}$ with $\mathbb{E}\mathrm{Q} = 0$ and $\mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+t) - \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}) \leq \delta_0$.

Proof. In the following, $\mu_{[a,b]}$ denotes the uniform distribution on the interval $[a,b]$. We write $y_0 := \frac{s_1+s_2}{2}$. Furthermore, if $y_0 = 0$, we define $\mathrm{Q} := \mu_{[-\varepsilon_0,\varepsilon_0]}$, and otherwise we define

$$\mathrm{Q} := \alpha\mu_{[-|\frac{y_0}{2}|,|\frac{y_0}{2}|]} + \frac{1-\alpha}{2}\mu_{[-y_0-\varepsilon_0,-y_0]} + \frac{1-\alpha}{2}\mu_{[y_0,y_0+\varepsilon_0]}\,,$$

where $\alpha \in (0,1)$ is a real number satisfying

$$\sup_{y\in[-|\frac{y_0}{2}|,|\frac{y_0}{2}|]} \Big|\frac{\psi(y-t)+\psi(y+t)}{2} - \psi(y)\Big| \leq \frac{\delta}{\alpha}\,.$$

Now we obviously have $\mathbb{E}\mathrm{Q} = 0$ in both cases. Moreover, if $y_0 \neq 0$, the construction together with Lemma 3.45 yields

$$\begin{aligned}
&\mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}_{L,\mathrm{Q}}(0)\\
&= \int_{\mathbb{R}} \frac{\psi(y-t)+\psi(y+t)}{2} - \psi(y)d\mathrm{Q}(y)\\
&= \alpha\int_{[-|\frac{y_0}{2}|,|\frac{y_0}{2}|]} \frac{\psi(y-t)+\psi(y+t)}{2} - \psi(y)\,d\mu_{[-|\frac{y_0}{2}|,|\frac{y_0}{2}|]}(y)\\
&\qquad +(1-\alpha)\int_{[y_0,y_0+\varepsilon_0]} \frac{\psi(y-t)+\psi(y+t)}{2} - \psi(y)\,d\mu_{[y_0,y_0+\varepsilon_0]}(y)\\
&\leq \delta + (1-\alpha)\int_{[0,\varepsilon_0]} \frac{\psi(s_1+\varepsilon)+\psi(s_2+\varepsilon)}{2} - \psi\Big(\frac{s_1+s_2}{2}+\varepsilon\Big)\,d\mu_{[0,\varepsilon_0]}(\varepsilon)\\
&\leq \delta_0 + \delta\,.
\end{aligned}$$

Furthermore, the case $y_0 = 0$ can be shown analogously, since $y_0 = 0$ implies $y - t = s_1 + y$ and $y + t = s_2 + y$. The last assertion follows if we repeat the construction above with $\alpha = 0$. □

Let us now establish our first two main results, which characterize losses L that are L_{mean}-calibrated with respect to $\mathcal{Q}^*_{\mathbb{R},\mathrm{sym}}(L)$ and $\mathcal{Q}^{(1)}_{\mathbb{R},\mathrm{sym}}$, respectively.

Theorem 3.48 (Mean calibration I). *Let $L : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ be a symmetric and continuous loss. Then the following statements are equivalent:*

i) L is L_{mean}-calibrated with respect to $\mathcal{Q}^*_{\mathbb{R},\text{sym}}(L)$.
ii) L is L_{mean}-calibrated with respect to $\mathcal{Q}^*_{\text{bounded,sym}}$.
iii) L is L_{mean}-calibrated with respect to $\mathcal{Q}^*_{[-M,M],\text{sym}}$ for all $M > 0$.
iv) L is convex, and its representing function ψ has its only minimum at 0.

Proof. *i)* $\Rightarrow$ *ii)* $\Rightarrow$ *iii)*. Trivial.

iv) $\Rightarrow$ *i)*. Assume that L is not L_{mean}-calibrated with respect to $\mathcal{Q}^*_{\mathbb{R},\text{sym}}(L)$. By Lemma 3.46, there then exist a $\mathrm{Q} \in \mathcal{Q}^*_{\mathbb{R},\text{sym}}(L)$ and a $t \neq 0$ with $\mathcal{C}_{L,\mathrm{Q}}(m+t) = \mathcal{C}^*_{L,\mathrm{Q}}$, where $m := \mathbb{E}\mathrm{Q}$. Using $\mathcal{C}_{L,\mathrm{Q}}(m) = \mathcal{C}^*_{L,\mathrm{Q}}$, which we know from Lemma 3.45, then yields

$$\int_{\mathbb{R}} \frac{\psi(y-t)+\psi(y+t)}{2} - \psi(y)\, d\mathrm{Q}^{(m)}(y) = \mathcal{C}_{L,\mathrm{Q}}(m+t) - \mathcal{C}_{L,\mathrm{Q}}(m) = 0\,,$$

and hence the convexity of ψ shows $\frac{\psi(y-m-t)+\psi(y-m+t)}{2} - \psi(y-m) = 0$ for Q-almost all $y \in \mathbb{R}$. The continuity of ψ and the assumption $\mathrm{Q}(m + [-\rho, \rho]) > 0$ for all $\rho > 0$, then guarantee that $\frac{\psi(y-m-t)+\psi(y-m+t)}{2} - \psi(y-m) = 0$ holds for $y := m$. However, by the symmetry of ψ, this implies $\psi(t) = \psi(0)$.

iii) $\Rightarrow$ *iv)*. Assume that ψ is not convex. Then Lemma A.6.17 shows that there exist $s_1, s_2 \in \mathbb{R}$ with $s_1 \neq s_2$ and $\frac{\psi(s_1)+\psi(s_2)}{2} - \psi(\frac{s_1+s_2}{2}) < 0$. By the continuity of ψ, we then find (3.56) for some suitable $\delta_0 < 0$ and $\varepsilon_0 > 0$, and hence Lemma 3.47 gives an $M > 0$, a $\mathrm{Q} \in \mathcal{Q}^*_{[-M,M],\text{sym}}$, and a $t^* \neq 0$ with $\mathcal{C}_{L,\mathrm{Q}}(t^*) < \mathcal{C}_{L,\mathrm{Q}}(0)$ and $\mathbb{E}\mathrm{Q} = 0$. Now observe that since ψ is continuous and Q has bounded support, the map $t \mapsto \mathcal{C}_{L,\mathrm{Q}}(t)$ is continuous on $\mathbb{R}$ by Lemma A.6.2. Let $(t_n) \subset \mathbb{R}$ be a sequence with $\mathcal{C}_{L,\mathrm{Q}}(t_n) \to \mathcal{C}^*_{L,\mathrm{Q}}$ for $n \to \infty$. Since our previous considerations showed $\mathcal{C}_{L,\mathrm{Q}}(0) \neq \mathcal{C}^*_{L,\mathrm{Q}}$, there must exist an $\varepsilon > 0$ and an $n_0 \in \mathbb{N}$ such that $|t_n| \geq \varepsilon$ for all $n \geq n_0$. Since $\mathbb{E}\mathrm{Q} = 0$, this shows

$$\delta_{\max,L_{\text{mean}},L}(\varepsilon, \mathrm{Q}) = \inf_{t \notin (-\varepsilon,\varepsilon)} \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}} \;\leq\; \mathcal{C}_{L,\mathrm{Q}}(t_n) - \mathcal{C}^*_{L,\mathrm{Q}}$$

for all $n \geq n_0$. For $n \to \infty$, we hence find $\delta_{\max}(\varepsilon, \mathrm{Q}) = 0$, and consequently L is convex. Finally, assume that there exists a $t \neq 0$ with $\psi(t) = \psi(0)$. Then we find $\mathcal{C}_{L,\mathrm{Q}}(t) = \mathcal{C}^*_{L,\mathrm{Q}}$ for the distribution Q defined by $\mathrm{Q}(\{0\}) = 1$, and hence we obtain $\delta_{\max}(|t|, \mathrm{Q}) = 0$. Therefore ψ has its only minimum at 0. □

Theorem 3.49 (Mean calibration II). *Let $L : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ be a symmetric and continuous loss. Then the following statements are equivalent:*

i) L is L_{mean}-calibrated with respect to $\mathcal{Q}^{(1)}_{\mathbb{R},\text{sym}}(L)$.
ii) L is L_{mean}-calibrated with respect to $\mathcal{Q}_{\text{bounded,sym}}$.
iii) L is L_{mean}-calibrated with respect to $\mathcal{Q}_{[-M,M],\text{sym}}$ for all $M > 0$.
iv) L is strictly convex.

Proof. *i)* $\Rightarrow$ *ii)* $\Rightarrow$ *iii)*. Trivial.

iv) $\Rightarrow$ *i)*. It immediately follows from Lemma 3.45 and Lemma 3.46.

iii) $\Rightarrow$ *iv)*. If L is L_{mean}-calibrated with respect to $\mathcal{Q}_{[-M,M],\text{sym}}$ for all $M > 0$, then Theorem 3.48 shows that L is convex. Let us suppose that its representing function ψ is *not* strictly convex. Then there are $r_1, r_2 \in \mathbb{R}$ with $r_1 \neq r_2$ and

$$\psi\Big(\frac{1}{2}r_1 + \frac{1}{2}r_2\Big) = \frac{1}{2}\psi(r_1) + \frac{1}{2}\psi(r_2)\,.$$

From this and Lemma A.6.17, we find (3.56) for $\delta_0 = 0$ and some suitable $s_1 \neq s_2$ and $\varepsilon_0 > 0$. Lemma 3.47 then gives an $M > 0$, a $\mathrm{Q} \in \mathcal{Q}_{[-M,M],\text{sym}}$, and a $t_0 \neq 0$, with $\mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q} + t_0) = \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q})$, and hence Lemma 3.46 shows that L is not L_{mean}-calibrated with respect to $\mathrm{Q} \in \mathcal{Q}_{[-M,M],\text{sym}}(L)$. □

Our next aim is to estimate the function $\varepsilon \mapsto \delta_{\max}(\varepsilon, \mathcal{Q})$ for some classes of distributions $\mathcal{Q} \subset \mathcal{Q}_{\mathbb{R},\text{sym}}$. To this end, we define the **modulus of convexity** of a function $f : I \to \mathbb{R}$ defined on some interval I by

$$\delta_f(\varepsilon) := \inf\left\{\frac{f(x_1) + f(x_2)}{2} - f\Big(\frac{x_1 + x_2}{2}\Big) : x_1, x_2 \in I \text{ with } |x_1 - x_2| \geq \varepsilon\right\},$$

where $\varepsilon > 0$. In addition we say that f is **uniformly convex** if $\delta_f(\varepsilon) > 0$ for all $\varepsilon > 0$. We refer to Section A.6.3 for some properties of the modulus of convexity and uniformly convex functions.

With the help of the modulus of convexity, we can now formulate the following theorem that estimates $\delta_{\max,L_{\text{mean}},L}(\varepsilon, \mathcal{Q})$ and characterizes uniform L_{mean}-calibration.

Theorem 3.50 (Uniform mean calibration). *Let L be a symmetric, convex loss with representing function ψ. Then the following statements are true:*

i) For all $M > 0$, $\varepsilon > 0$, and $\mathcal{Q}^*_{[-M,M],\text{sym}} \subset \mathcal{Q} \subset \mathcal{Q}_{[-M,M],\text{sym}}$, we have

$$\delta_{\psi_{|[-(2M+\varepsilon),2M+\varepsilon]}}(2\varepsilon) \;\leq\; \delta_{\max,L_{\text{mean}},L}(\varepsilon, \mathcal{Q}) \;\leq\; \delta_{\psi_{|[-M/2,M/2]}}(2\varepsilon)\,. \quad (3.57)$$

Moreover, the following statements are equivalent:

a) L is uniformly L_{mean}-calibrated w.r.t. $\mathcal{Q}^*_{[-M,M],\text{sym}}$ for all $M > 0$.
b) L is uniformly L_{mean}-calibrated w.r.t. $\mathcal{Q}_{[-M,M],\text{sym}}$ for all $M > 0$.
c) The function ψ is strictly convex.

ii) For all $\varepsilon > 0$, we have

$$\delta_\psi(2\varepsilon) = \delta_{\max,L_{\text{mean}},L}(\varepsilon, \mathcal{Q}_{\mathbb{R},\text{sym}}(L)) = \delta_{\max}(\varepsilon, \mathcal{Q}^*_{\text{bounded,sym}})\,. \quad (3.58)$$

Moreover, the following statements are equivalent:

a) L is uniformly L_{mean}-calibrated with respect to $\mathcal{Q}^{(1)}_{\mathbb{R},\text{sym}}(L)$.
b) L is uniformly L_{mean}-calibrated with respect to $\mathcal{Q}^*_{\text{bounded,sym}}$.
c) The function ψ is uniformly convex.

Proof. i). Let $\mathrm{Q} \in \mathcal{Q}_{[-M,M],\mathrm{sym}}$. Then we have $\mathbb{E}\mathrm{Q} \in [-M, M]$, and hence Lemmas 3.45 and 3.46 yield

$$\begin{aligned}\delta_{\max}(\varepsilon, \mathrm{Q}) &= \int_{[-M,M]} \frac{\psi\big(y - \mathbb{E}\mathrm{Q} - \varepsilon\big) + \psi\big(y - \mathbb{E}\mathrm{Q} + \varepsilon\big)}{2} - \psi(y - \mathbb{E}\mathrm{Q})\, d\mathrm{Q}(y) \\ &\geq \delta_{\psi_{|[-(2M+\varepsilon),2M+\varepsilon]}}(2\varepsilon).\end{aligned}$$

This shows the first inequality of (3.57). To prove the second inequality, we observe that it suffices to consider the case $\varepsilon \leq M/2$ since for $\varepsilon > M/2$ we have $\delta_{\psi_{|[-M/2,M/2]}}(2\varepsilon) = \infty$. Let us now fix an $n \geq 1$. Then there exist $s_1, s_2 \in [-M/2, M/2]$ with $s_1 - s_2 \geq 2\varepsilon$ and

$$\frac{\psi(s_1) + \psi(s_2)}{2} - \psi\left(\frac{s_1 + s_2}{2}\right) < \delta_{\psi_{|[-M/2,M/2]}}(2\varepsilon) + \frac{1}{n} =: \delta_0 < \infty\,.$$

By the continuity of ψ, there thus exists an $\varepsilon_0 \in (0, M/2]$ such that (3.56) is satisfied for δ_0, and consequently Lemma 3.47 gives a $\mathrm{Q} \in \mathcal{Q}^*_{[-M,M],\mathrm{sym}}$ with

$$\mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q} + t) - \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}) \leq \delta_{\psi_{|[-M/2,M/2]}}(2\varepsilon) + \frac{2}{n}\,,$$

where $t := \frac{s_1 - s_2}{2}$. Using $t \geq \varepsilon$ and Lemma 3.46, we hence find

$$\delta_{\max}(\varepsilon, \mathcal{Q}^*_{[-M,M],\mathrm{sym}}) \leq \delta_{\max}(\varepsilon, \mathrm{Q}) \leq \delta_{\psi_{|[-M/2,M/2]}}(2\varepsilon) + \frac{2}{n}.$$

Since this holds for all $n \geq 1$, the second inequality of (3.57) follows. Finally, from Lemma A.6.17, we know that ψ is strictly convex if and only if $\delta_{\psi_{|[-B,B]}}(\varepsilon) > 0$ for all B and $\varepsilon > 0$, and hence the characterization follows.

ii). Analogously to the proof of the first inequality in (3.57), we find

$$\delta_\psi(2\varepsilon) \leq \delta_{\max}(\varepsilon, \mathcal{Q}_{\mathbb{R},\mathrm{sym}}(L))\,, \qquad \varepsilon > 0.$$

Furthermore, analogously to the proof of the second inequality in (3.57), we obtain

$$\delta_{\max}(\varepsilon, \mathcal{Q}^*_{\mathrm{bounded,sym}}) \leq \delta_\psi(2\varepsilon)\,, \qquad \varepsilon > 0,$$

and hence (3.58) is proved. Finally, the characterization is a trivial consequence of (3.58). □

The preceding theorem shows that the modulus of convexity completely determines whether a symmetric loss is uniformly L_{mean}-calibrated with respect to $\mathcal{Q}^{(1)}_{\mathbb{R},\mathrm{sym}}(L)$ or $\mathcal{Q}^*_{\mathrm{bounded,sym}}$. Unfortunately, Lemma A.6.19 shows that, for all distance-based losses of upper growth type $p < 2$, we have $\delta_\psi(\varepsilon) = 0$ for *all* $\varepsilon > 0$. In particular, Lipschitz continuous, distance-based losses, which are of special interest for robust regression methods (see Chapter 10), are *not* uniformly calibrated with respect to $\mathcal{Q}^{(1)}_{\mathbb{R},\mathrm{sym}}(L)$ or $\mathcal{Q}^*_{\mathrm{bounded,sym}}$, and consequently we cannot establish *distribution independent* relations between the

Table 3.3. Some symmetric loss functions and corresponding upper and lower bounds for the moduli of convexity $\delta_{\psi_{|[-B,B]}}(2\varepsilon)$, $0 < \varepsilon \leq B$. The asymptotics for the L_p-loss, $1 < p < 2$, are computed in Exercise 3.12. For the L_p-loss, $p \geq 2$, and Huber's loss, the lower bounds can be found by Clarkson's inequality (see Lemma A.5.24), and the upper bounds can be found by picking suitable $t_1, t_2 \in [-B, B]$. The calculations for the logistic loss can be found in Example 3.51.

Loss Function	Lower Bound of $\delta_{\psi_{\|[-B,B]}}(2\varepsilon)$	Upper Bound of $\delta_{\psi_{\|[-B,B]}}(2\varepsilon)$
$L_{\text{1-dist}}$	0	0
$L_{p\text{-dist}}$, $p \in (1,2)$	$\frac{p(p-1)}{2} B^{p-2} \varepsilon^2$	$\frac{p}{2(p-1)^2} B^{p-2} \varepsilon^2$
$L_{p\text{-dist}}$, $p \in [2,\infty)$	ε^p	ε^p
$L_{\text{r-logist}}$	$\frac{1-e^{-\varepsilon}}{2} \ln \frac{e^B+e^{2\varepsilon}}{e^B+e^{\varepsilon}}$	$(1-e^{-\varepsilon}) \ln \frac{e^B+e^{2\varepsilon}}{e^B+e^{\varepsilon}}$
$L_{\alpha\text{-Huber}}$, $\alpha > 0$	$\frac{\varepsilon^2}{2}$ if $B \leq \alpha$ 0 else	$\frac{\varepsilon^2}{2}$ if $B \leq \alpha$ 0 else

excess L-risks and $\mathcal{R}_{L_{\text{mean}},\text{P}}(\,\cdot\,)$ in the sense of Question 3.2. On the other hand, symmetric, strictly convex losses L are L_{mean}-calibrated with respect to $\mathcal{Q}^{(1)}_{\mathbb{R},\text{sym}}(L)$, and hence we can show analogously to Theorem 3.61 below that $f_n \to \mathbb{E}(Y|\,\cdot\,)$ in probability P_X whenever $\mathcal{R}_{L,\text{P}}(f_n) \to \mathcal{R}^*_{L,\text{P}}$ and P is of type $\mathcal{Q}^{(1)}_{\mathbb{R},\text{sym}}(L)$. In addition, if we restrict our considerations to $\mathcal{Q}_{[-M,M],\text{sym}}$ or $\mathcal{Q}^*_{[-M,M],\text{sym}}$, then every symmetric, strictly convex loss becomes uniformly L_{mean}-calibrated, and in this case $\delta_{\psi_{|[-B,B]}}(\,\cdot\,)$, $B > 0$, can be used to describe the corresponding calibration function. For some important losses, we have listed the behavior of $\delta_{\psi_{|[-B,B]}}(\,\cdot\,)$ in Table 3.3. Furthermore, Lemma A.6.19 establishes a formula for the modulus of convexity that often helps to bound the modulus. The following example illustrates this.

Example 3.51. Recall from Example 2.40 that the **logistic loss for regression** is the symmetric loss represented by $\psi(t) := -\ln \frac{4e^t}{(1+e^t)^2}$, $t \in \mathbb{R}$. Let us show, that for $B > 0$ and $\varepsilon \in (0, B]$, we have

$$\frac{1-e^{-\varepsilon}}{2} \ln \frac{e^B+e^{2\varepsilon}}{e^B+e^{\varepsilon}} \;\leq\; \delta_{\psi_{|[-B,B]}}(2\varepsilon) \;\leq\; (1-e^{-\varepsilon}) \ln \frac{e^B+e^{2\varepsilon}}{e^B+e^{\varepsilon}}\,.$$

To see this, we first observe that $\psi'(t) = \frac{e^t-1}{e^t+1}$ for all $t \in \mathbb{R}$, and hence we obtain

$$\psi'(t) - \psi'(t-\varepsilon) = \frac{e^t-1}{e^t+1} - \frac{e^t-e^{\varepsilon}}{e^t+e^{\varepsilon}} = \frac{2e^t(e^{\varepsilon}-1)}{(e^t+1)(e^t+e^{\varepsilon})}$$

for all $t \in \mathbb{R}$ and $\varepsilon \geq 0$. Consequently, we have

$$\frac{e^\varepsilon - 1}{e^t + e^\varepsilon} \;\le\; \psi'(t) - \psi'(t-\varepsilon) \;\le\; 2\frac{e^\varepsilon - 1}{e^t + e^\varepsilon}$$

for all $t \ge 0$ and $\varepsilon \ge 0$. Furthermore, for $\varepsilon > 0$ an easy calculation gives

$$\begin{aligned}\inf_{x\in[0,B-\varepsilon]} \int_x^{x+\varepsilon} \frac{e^\varepsilon - 1}{e^t + e^\varepsilon} dt &= \int_{B-\varepsilon}^{B} \frac{e^\varepsilon - 1}{e^t + e^\varepsilon} dt = (1 - e^{-\varepsilon}) \Big(t - \ln\big(e^t + e^\varepsilon\big) \Big) \Big|_{t=B-\varepsilon}^{B} \\ &= (1 - e^{-\varepsilon}) \ln \frac{e^B + e^{2\varepsilon}}{e^B + e^\varepsilon} \,.\end{aligned}$$

Using Lemma A.6.19 then yields the assertion. ◁

In Theorem 3.48, we have seen that for $\mathrm{Q} \in \mathcal{Q}^*_{\mathbb{R},\mathrm{sym}}(L)$ we may have $\delta_{\max}(\varepsilon, \mathrm{Q}) > 0$, $\varepsilon > 0$, even if L is not strictly convex. The key reason for this possibility was the assumption that Q has some mass around its center. Now recall that in the proof of the upper bounds of Theorem 3.50 we used the fact that for general $\mathrm{Q} \in \mathcal{Q}^*_{\mathbb{R},\mathrm{sym}}$ this mass can be arbitrarily small. However, if we enforce lower bounds on this mass, the construction of this proof no longer works. Instead, it turns out that we can establish lower bounds on $\delta_{\max}(\varepsilon, \mathrm{Q})$, as the following example illustrates (see also Example 3.67).

Example 3.52. Recall that the **absolute distance loss** is the symmetric loss represented by $\psi(t) = |t|$, $t \in \mathbb{R}$. Then, for all $\mathrm{Q} \in \mathcal{Q}^{(1)}_{\mathbb{R},\mathrm{sym}}$ and $\varepsilon > 0$, we have

$$\delta_{\max, L_{\mathrm{mean}}, L_{1\text{-dist}}}(\varepsilon, \mathrm{Q}) = \int_0^\varepsilon \mathrm{Q}^{(\mathbb{E}\mathrm{Q})}\big((-s, s)\big)\, ds \,. \tag{3.59}$$

To see this, recall that for symmetric distributions the mean equals the median, i.e., the 1/2-quantile. Now (3.59) follows from Proposition 3.9. ◁

The results in this section showed that using symmetric surrogate losses for regression problems requires some care: for example, let us suppose that the primary goal of the regression problem is to estimate the conditional mean. If we only know that the conditional distributions $\mathrm{P}(\,\cdot\,|x)$, $x \in X$, have finite variances (and expect these distributions to be rather asymmetric), then the least squares loss is the only reasonable, distance-based choice by Proposition 3.44. However, if we know that these distributions are (almost) symmetric, then symmetric, strictly convex, and Lipschitz continuous losses such as the logistic loss can be reasonable alternatives. In addition, if we are confident that these conditional distributions are also rather concentrated around their mean, e.g., in the form of $\mathrm{Q}^{(\mathbb{E}\mathrm{Q})}((-s, s)) > c_{\mathrm{Q}} s^q$ for small $s > 0$, then even the absolute distance loss can be a good choice. Finally, if we additionally expect that the data set contains extreme outliers, then the logistic loss or the absolute distance loss may actually be a better choice than the least squares loss. However, recall that such a decision only makes sense if the noise distribution is (almost) symmetric.

3.8 Surrogate Losses for the Density Level Problem

In this section, our goal is to find supervised loss functions that are calibrated with respect to the density level detection loss L_{DLD} introduced in Example 2.9. To this end, let us first recall that in the density level detection scenario our learning goal was to identify the ρ-level set $\{g > \rho\}$ of an unknown density $g : X \to [0, \infty)$ whose reference distribution μ on X is known. Unfortunately, the loss L_{DLD} formalizing this learning goal does depend on the unknown density g, and thus we cannot compute its associated risks. Consequently, our goal in this section is to find *supervised* surrogates for L_{DLD} that do not depend on g. At first glance, this goal seems to be rather impossible since supervised losses require labels that do not exist in the description of the DLD learning scenario. Therefore, our first goal is to resolve this issue by introducing *artificial* labels. To this end, we need the following definition.

Definition 3.53. *Let μ be a distribution on some X and $Y := \{-1, 1\}$. Furthermore, let $g : X \to [0, \infty)$ be measurable with $\|g\|_{L_1(\mu)} = 1$. Then, for $\rho > 0$, we write $g\mu \ominus_\rho \mu$ for the distribution* P *on $X \times Y$ that is defined by*

$$\begin{aligned} \mathrm{P}_X &:= \frac{g + \rho}{1 + \rho}\,\mu\,, \\ \mathrm{P}(y = 1|x) &:= \frac{g(x)}{g(x) + \rho}\,, \qquad x \in X. \end{aligned}$$

An elementary calculation shows that for measurable $A \subset X \times Y$ we have

$$g\mu \ominus_\rho \mu\,(A) = \frac{1}{1 + \rho}\mathbb{E}_{x\sim g\mu}\mathbf{1}_A(x, 1) + \frac{\rho}{1 + \rho}\mathbb{E}_{x\sim\mu}\mathbf{1}_A(x, -1)\,, \tag{3.60}$$

and hence $\mathrm{P} := g\mu \ominus_\rho \mu$ describes a binary classification problem in which the negative samples are drawn from the distribution μ with probability $\frac{\rho}{1+\rho}$ and in which the positive samples are drawn from the distribution $g\mu$ with probability $\frac{1}{1+\rho}$.

We have already mentioned in Example 2.9 that we are primarily interested in the quantity $\mathcal{R}_{L_{\mathrm{DLD}},\mu}(f)$, which describes the discrepancy of the estimated level set $\{f \geq 0\}$ to the true ρ-level set. Now observe that, for $\mathrm{P} := g\mu \ominus_\rho \mu$ and measurable $f : X \to \mathbb{R}$, we have

$$\mathcal{R}_{L_{\mathrm{DLD}},\mu}(f) = \int_X L_{\mathrm{DLD}}(x, f(x))\,d\mu(x) = \int_X L_{\mathrm{DLD}}(x, f(x))\frac{1 + \rho}{g(x) + \rho}\,d\mathrm{P}_X(x)\,,$$

and consequently we can describe the DLD learning scenario by P and the detection loss $\bar{L} : X \times \mathbb{R} \to [0, \infty)$ defined by

$$\bar{L}(x, t) := L_{\mathrm{DLD}}(x, t)\frac{1 + \rho}{g(x) + \rho}\,, \qquad x \in X,\ t \in \mathbb{R}. \tag{3.61}$$

The first benefit of this reformulation is that our new target risk $\mathcal{R}_{\bar{L},\mathrm{P}}(\,\cdot\,)$ is defined by a distribution P, which produces labels, and consequently it makes

sense to look for supervised surrogates for $\bar{L}$. Furthermore, we have access to P via (3.60) in the sense that *a)* the distribution $g\mu$ can be estimated from the unlabeled samples given in the DLD scenario, see Example 2.9, and *b)* both μ and ρ are *known*. This makes it possible to construct an empirical approximation of P that can then lead to learning algorithms based on this approximation. For some literature in this direction, we refer to Section 2.5 and to the end of Section 8.6. The second benefit of considering the $\bar{L}$-risk is that P describes a *classification problem*, and hence it seems natural to *a)* investigate $\bar{L}$-calibration with the help of classification calibration and *b)* use classification algorithms for the DLD learning scenario. In order to confirm this intuition, let us consider the function $\bar{L}_{\mathrm{DLD}} : [0,1] \times \mathbb{R} \to [0,\infty)$ defined by

$$\bar{L}_{\mathrm{DLD}}(\eta, t) := (1-\eta)\mathbf{1}_{(-\infty,0)}\big((2\eta-1)\operatorname{sign} t\big)\,. \tag{3.62}$$

Using the identification $\eta = \mathrm{Q}(\{1\})$ between $\eta \in [0,1]$ and $\mathrm{Q} \in \mathcal{Q}_Y$, where $Y := \{-1,1\}$, we can regard the function $\bar{L}_{\mathrm{DLD}}$ as a template loss. For $\mathrm{P} = g\mu \ominus_\rho \mu$, the P-instance $\bar{L}_{\mathrm{DLD,P}}$ of $\bar{L}_{\mathrm{DLD}}$ then becomes

$$\begin{aligned}
\bar{L}_{\mathrm{DLD,P}}(x,t) = \bar{L}_{\mathrm{DLD}}\big(\mathrm{P}(\,\cdot\,|x), t\big) &= (1-\eta(x))\mathbf{1}_{(-\infty,0)}\big((2\eta(x)-1)\operatorname{sign} t\big) \\
&= \frac{\rho}{g(x)+\rho}\mathbf{1}_{(-\infty,0)}\big((g(x)-\rho)\operatorname{sign} t\big) \\
&= \frac{\rho}{1+\rho}\bar{L}(x,t)\,,
\end{aligned}$$

where we used $\eta(x) := \mathrm{P}(y=1|x) = \frac{g(x)}{g(x)+\rho}$. In other words, the P-instance $\bar{L}_{\mathrm{DLD,P}}$ of $\bar{L}_{\mathrm{DLD}}$ equals our detection loss $\bar{L}$ up to the constant $\frac{\rho}{1+\rho}$, and hence we obtain

$$\mathcal{R}_{L_{\mathrm{DLD}},\mu}(f) = \mathcal{R}_{\bar{L},\mathrm{P}}(f) = \frac{1+\rho}{\rho}\mathcal{R}_{\bar{L}_{\mathrm{DLD,P}},\mathrm{P}}(f) \tag{3.63}$$

for $\mathrm{P} = g\mu \ominus_\rho \mu$ and all measurable functions $f : X \to \mathbb{R}$. Consequently, suitable supervised surrogates for the DLD problem are exactly the losses that are $\bar{L}_{\mathrm{DLD}}$-calibrated in the following sense.

Definition 3.54. *Let $Y := \{-1,1\}$ and $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss. We say that L is **(uniformly) density level detection calibrated** if L is (uniformly) $\bar{L}_{\mathrm{DLD}}$-calibrated with respect to $\mathcal{Q}_Y$.*

In order to identify DLD-calibrated losses, we need to know the corresponding calibration function. This function is computed in the next lemma.

Lemma 3.55 (Calibration function for DLD). *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss function. Then, for all $\eta \in [0,1]$ and $\varepsilon \in (0,\infty]$, we have*

$$\delta_{\max,\bar{L}_{\mathrm{DLD}},L}(\varepsilon,\eta) = \begin{cases} \infty & \text{if } \varepsilon > 1-\eta \\ \inf_{t\in\mathbb{R}:(2\eta-1)\operatorname{sign} t<0} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} & \text{if } \varepsilon \le 1-\eta. \end{cases}$$

Proof. A simple calculation shows $\mathcal{C}^*_{\bar{L}_{\mathrm{DLD}},\eta} = 0$, and consequently we obtain $\mathcal{M}_{\bar{L}_{\mathrm{DLD}},\eta}(\varepsilon) = \mathbb{R}$ if $\varepsilon > 1-\eta$, and $\mathcal{M}_{\bar{L}_{\mathrm{DLD}},\eta}(\varepsilon) = \{t \in \mathbb{R} : (2\eta-1)\operatorname{sign} t \geq 0\}$ otherwise. From this we immediately find the assertion. □

With the help of the preceding lemma, we now obtain the first main result, which compares classification calibration with $\bar{L}_{\mathrm{DLD}}$-calibration.

Theorem 3.56 (DLD-calibration). *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss and $\eta \in [0,1]$. Then, for all $0 \leq \varepsilon \leq \min\{1-\eta, |2\eta-1|\}$, we have*

$$\delta_{\max,\bar{L}_{\mathrm{DLD}},L}(\varepsilon,\eta) \geq \delta_{\max,L_{\mathrm{class}},L}(\varepsilon,\eta)\,,$$

and consequently L is DLD-calibrated if L is classification calibrated. Moreover, if L is continuous, then the inequality above becomes an equality and L is classification calibrated if and only if L is DLD-calibrated.

Proof. Combining Lemma 3.55 with Lemma 3.32 yields

$$\begin{aligned}\delta_{\max,\bar{L}_{\mathrm{DLD}},L}(\varepsilon,\eta) = \inf_{\substack{t\in\mathbb{R}:\\ (2\eta-1)\operatorname{sign} t<0}} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} \;&\geq \inf_{\substack{t\in\mathbb{R}:\\ (2\eta-1)\operatorname{sign} t\leq 0}} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} \\ &= \delta_{\max,L_{\mathrm{class}},L}(\varepsilon,\eta)\,.\end{aligned}$$

Now assume that L is continuous. Since there is nothing to prove for $\eta = 1/2$, we additionally assume $\eta \neq 1/2$. Then the assertion can be found by using the continuity of $t \mapsto \mathcal{C}_{L,\eta}(t)$ in the estimate above. □

By the results on classification calibrated, margin-based losses from Section 3.4, we immediately obtain a variety of DLD-calibrated losses. Furthermore, the P-instances of $\bar{L}_{\mathrm{DLD}}$ are bounded and hence Theorem 3.27 yields

$$\mathcal{R}_{L,\mathrm{P}}(f_n) \to \mathcal{R}^*_{L,\mathrm{P}} \qquad \Longrightarrow \qquad \mathcal{R}_{L_{\mathrm{DLD}},\mu}(f_n) \to 0$$

whenever $\mathrm{P} = g\mu \ominus_\rho \mu$ and L is classification calibrated. In addition, one can show that for $L := L_{\mathrm{class}}$ the converse implication is also true. For details, we refer to Exercise 3.13.

Our next goal is to identify *uniformly* DLD-calibrated losses. The following theorem gives a complete, though rather disappointing, solution.

Theorem 3.57 (No uniform DLD-calibration). *There exists no supervised loss $L : Y \times \mathbb{R} \to [0,\infty)$ that is uniformly $\bar{L}_{\mathrm{DLD}}$-calibrated with respect to both $\{\mathrm{Q} \in \mathcal{Q}_Y : \mathrm{Q}(\{1\}) \in [0,1/2)\}$ and $\{\mathrm{Q} \in \mathcal{Q}_Y : \mathrm{Q}(\{1\}) \in (1/2,3/4]\}$. In particular, there exists no uniform DLD-calibrated supervised loss.*

Proof. Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss. For $\eta \in [0,1]$, we define

$$g^+(\eta) = \inf_{t<0} \mathcal{C}_{L,\eta}(t) \qquad \text{and} \qquad g^-(\eta) = \inf_{t\geq 0} \mathcal{C}_{L,\eta}(t)\,.$$

Then the functions $g^+ : [0,1] \to [0,\infty)$ and $g^- : [0,1] \to [0,\infty)$ can be defined by suprema taken over affine linear functions in $\eta \in \mathbb{R}$, and since g^+ and g^-

are also finite for $\eta \in [0,1]$, we find by Lemma A.6.4 that g^+ and g^- are continuous at every $\eta \in [0,1]$. Moreover, we have $\mathcal{C}^*_{L,\eta} = \min\{g^+(\eta), g^-(\eta)\}$ for all $\eta \in [0,1]$, and hence $\mathcal{C}^*_{L,\eta}$ is continuous in η. Let us first consider the case $\mathcal{C}^*_{L,1/2} = g^+(1/2)$. To this end, we first observe that there exists a sequence $(t_n) \subset (-\infty, 0)$ with

$$g^+\bigl(1/2 + 1/n\bigr) \;\leq\; \mathcal{C}_{L,1/2+1/n}(t_n) \;\leq\; g^+\bigl(1/2 + 1/n\bigr) + 1/n \tag{3.64}$$

for all $n \geq 1$. Moreover, our assumption $\mathcal{C}^*_{L,1/2} = g^+(1/2)$ yields

$$\begin{aligned}\bigl|\mathcal{C}_{L,1/2+1/n}(t_n) - \mathcal{C}^*_{L,1/2+1/n}\bigr| \leq{}& \bigl|\mathcal{C}_{L,1/2+1/n}(t_n) - g^+(1/2+1/n)\bigr| \\ &+ \bigl|g^+\bigl(1/2+1/n\bigr) - g^+(1/2)\bigr| \\ &+ \bigl|\mathcal{C}^*_{L,1/2} - \mathcal{C}^*_{L,1/2+1/n}\bigr|\end{aligned}$$

for all $n \geq 1$. By (3.64) and the continuity of g^+ and $\eta \mapsto \mathcal{C}^*_{L,\eta}$, we hence find

$$\lim_{n\to\infty} \bigl|\mathcal{C}_{L,1/2+1/n}(t_n) - \mathcal{C}^*_{L,1/2+1/n}\bigr| = 0\,.$$

For $\mathcal{Q} := \bigl\{\mathrm{Q} \in \mathcal{Q}_Y : \mathrm{Q}(\{1\}) \in (1/2, 3/4]\bigr\}$, Lemma 3.55, the definition g^+, and (3.64) then yield

$$\begin{aligned}\delta_{\max,\bar{L}_{\mathrm{DLD}},L}(\varepsilon,\mathcal{Q}) = \inf_{\eta\in(\frac{1}{2},\frac{3}{4}]} g^+(\eta) - \mathcal{C}^*_{L,\eta} &\leq \inf_{n\geq 1} \mathcal{C}_{L,1/2+1/n}(t_n) - \mathcal{C}^*_{L,1/2+1/n} \\ &= 0\,.\end{aligned}$$

Consequently, L is not uniformly $\bar{L}_{\mathrm{DLD}}$-calibrated with respect to $\mathcal{Q}$. Finally, in the case $\mathcal{C}^*_{L,1/2} = g^-(1/2)$, we can analogously show that L is not uniformly $\bar{L}_{\mathrm{DLD}}$-calibrated with respect to $\{\mathrm{Q} \in \mathcal{Q}_Y : \mathrm{Q}(\{1\}) \in [0, 1/2)\}$. □

The preceding theorem shows that there exists no uniformly DLD-calibrated, supervised loss. Now recall that Theorem 3.24 showed that uniform calibration is *necessary* to establish inequalities between excess risks if essentially no assumptions on the data-generating distribution are imposed.[1] Together with Theorem 3.57, we consequently see that it is *impossible* to find a supervised loss $L : Y \times \mathbb{R} \to [0,\infty)$ and an increasing function $\delta : [0,\infty] \to [0,\infty]$ such that $\delta(0) = 0$, $\delta(\varepsilon) > 0$ for all $\varepsilon > 0$, and

$$\delta\bigl(\mathcal{R}_{L_{\mathrm{DLD}},\mu}(f)\bigr) \;\leq\; \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \tag{3.65}$$

for *all* μ, g, ρ, f, and $\mathrm{P} := g\mu \ominus_\rho \mu$. However, in the DLD learning scenario, we actually *know* μ and ρ, and hence the question remains whether for certain *fixed* μ and ρ there exists a non-trivial function δ satisfying (3.65). Unfortunately, Steinwart (2007) showed that the answer is again no.

[1] Formally, the result only holds for loss functions and *not template losses*. However, it is quite straightforward to see that the proof of Theorem 3.24 can be easily modified to establish an analogous result for instances of template losses.

3.9 Self-Calibrated Loss Functions

Given a loss L and a distribution P such that an *exact* minimizer $f^*_{L,\mathrm{P}}$ of $\mathcal{R}_{L,\mathrm{P}}(\,\cdot\,)$ exists, one may ask whether, and in which sense, approximate minimizers f of $\mathcal{R}_{L,\mathrm{P}}(\,\cdot\,)$ approximate $f^*_{L,\mathrm{P}}$. For example, in binary classification, one often wants to find a decision function f that not only has a small classification error but also estimates the conditional probability $\mathrm{P}(y=1|x)$. Now assume that we have found an f whose excess L-risk is small for a suitable surrogate L of the classification loss (recall Section 3.4 for examples of such surrogates). Assume further that the L-risk has a *unique* minimizer $f^*_{L,\mathrm{P}}$ that, in addition, has a one-to-one correspondence to the conditional probability. If we have a positive answer to the question above, we can then use a suitable transformation of $f(x)$ to estimate $\mathrm{P}(y=1|x)$. An important example of such a loss, namely the logistic loss for classification, is discussed in Example 3.66. Moreover, we will discuss how the pinball loss can be used to estimate quantiles. The main goal of this section is, however, to provide some general answers to the question above.

Let us begin by introducing some notation. To this end, let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss function for some $Y \subset \mathbb{R}$ closed. We write

$$\begin{aligned}\mathcal{Q}_{\min}(L) &:= \bigl\{\mathrm{Q} : \mathrm{Q} \text{ is a distribution on } Y \text{ with } \mathcal{M}_{L,\mathrm{Q}}(0^+) \neq \emptyset\bigr\}\,,\\ \mathcal{Q}_{1\text{-}\min}(L) &:= \bigl\{\mathrm{Q} \in \mathcal{Q}_{\min}(L) : \exists\, t^*_{L,\mathrm{Q}} \in \mathbb{R} \text{ such that } \mathcal{M}_{L,\mathrm{Q}}(0^+) = \{t^*_{L,\mathrm{Q}}\}\bigr\}\,,\end{aligned}$$

i.e., $\mathcal{Q}_{\min}(L)$ contains the distributions on Y whose inner L-risks have an exact minimizer, while $\mathcal{Q}_{1\text{-}\min}(L)$ contains the distributions on Y whose inner L-risks have exactly one exact minimizer. Obviously, $\mathcal{Q}_{1\text{-}\min}(L) \subset \mathcal{Q}_{\min}(L)$ holds, and for strictly convex losses L, both sets actually coincide. Moreover, note that by Lemma 3.10 we have $\mathcal{C}^*_{L,\mathrm{Q}} < \infty$ for all $\mathrm{Q} \in \mathcal{Q}_{\min}(L)$. For $\mathrm{Q} \in \mathcal{Q}_{\min}(L)$, we now define the **self-calibration loss** of L by

$$\breve{L}(\mathrm{Q},t) \;:=\; \operatorname{dist}\bigl(t, \mathcal{M}_{L,\mathrm{Q}}(0^+)\bigr) := \inf_{t' \in \mathcal{M}_{L,\mathrm{Q}}(0^+)} |t-t'|\,, \qquad t \in \mathbb{R}, \tag{3.66}$$

i.e., $\breve{L}(\mathrm{Q},t)$ measures the distance of t to the set of elements minimizing $\mathcal{C}_{L,\mathrm{Q}}$. The next lemma shows that the self-calibration loss is a template loss.

Lemma 3.58. *Let $Y \subset \mathbb{R}$ be closed and $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss. Then $\breve{L} : \mathcal{Q}_{\min}(L) \times \mathbb{R} \to [0,\infty)$ defined by (3.66) is a template loss.*

Proof. Let X be a complete measurable space and P be a distribution on $X \times Y$ with $\mathrm{P}(\,\cdot\,|x) \in \mathcal{Q}_{\min}(L)$ for all $x \in X$. We write $\bar{X} := X \times \mathbb{R}$ and $Z := \mathbb{R}$. Furthermore, for $\bar{x} = (x,t) \in \bar{X}$ and $t' \in Z$, we define

$$\begin{aligned}h(\bar{x},t') &:= \mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x)}(t') - \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x)}\,,\\ F(\bar{x}) &:= \{t^* \in \mathbb{R} : h(\bar{x},t^*) = 0\}\,,\end{aligned}$$

and $\varphi(\bar{x},t') := |t-t'|$. For the P-instance $\breve{L}_{\mathrm{P}}$ of $\breve{L}$, we then have

$$\breve{L}_{\mathrm{P}}(x,t) = \inf_{t' \in \mathcal{M}_{L,\mathrm{P}(\,\cdot\,|x)}(0^+)} |t-t'| = \inf_{t' \in F(\bar{x})} \varphi(\bar{x},t')\,,$$

and consequently we obtain the assertion by part *iii)* of Lemma A.3.18. □

It is almost needless to say that the main statement of the preceding lemma is the *measurability* of the instances of $\breve{L}$. Now note that the definition of $\breve{L}$ immediately gives $\mathcal{C}^*_{\breve{L},\mathrm{Q}} = 0$, and therefore we have

$$\mathcal{M}_{\breve{L},\mathrm{Q}}(\varepsilon) = \bigl\{t \in \mathbb{R} : \breve{L}(\mathrm{Q},t) < \varepsilon\bigr\} = \bigl\{t \in \mathbb{R} \colon \exists t' \in \mathcal{M}_{L,\mathrm{Q}}(0^+) \text{ with } |t-t'| < \varepsilon\bigr\}$$

for all $\mathrm{Q} \in \mathcal{Q}_{\min}(L)$ and $\varepsilon \in [0,\infty]$. Moreover, we have already mentioned in Section 3.6 that the results of Lemma 3.14 remain true for template losses. By (3.16), the **self-calibration function** $\delta_{\max,\breve{L},L}(\,\cdot\,,\mathrm{Q})$, which can be computed by

$$\delta_{\max,\breve{L},L}(\varepsilon,\mathrm{Q}) \;=\; \inf_{\substack{t\in\mathbb{R}\\ \operatorname{dist}(t,\mathcal{M}_{L,\mathrm{Q}}(0^+))\geq\varepsilon}} \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}} \tag{3.67}$$

for all $\varepsilon \in [0,\infty]$, thus satisfies

$$\delta_{\max,\breve{L},L}\bigl(\operatorname{dist}(t,\mathcal{M}_{L,\mathrm{Q}}(0^+)),\mathrm{Q}\bigr) \;\leq\; \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}}\,, \qquad t \in \mathbb{R},$$

for all $\mathrm{Q} \in \mathcal{Q}_{\min}(L)$. Note that for $\mathrm{Q} \in \mathcal{Q}_{1\text{-}\min}(L)$ this inequality becomes

$$\delta_{\max,\breve{L},L}\bigl(|t-t^*_{L,\mathrm{Q}}|,\mathrm{Q}\bigr) \;\leq\; \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}}\,, \qquad t \in \mathbb{R},$$

where $\mathcal{M}_{L,\mathrm{Q}}(0^+) = \{t^*_{L,\mathrm{Q}}\}$. Consequently, the self-calibration function indeed quantifies how well an approximate $\mathcal{C}_{L,\mathrm{Q}}$-minimizer t approximates the exact minimizer $t^*_{L,\mathrm{Q}}$. This motivates the following, main definition of this section.

Definition 3.59. *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss function and $\mathcal{Q} \subset \mathcal{Q}_{\min}(L)$. We say that L is **(uniformly) self-calibrated** with respect to $\mathcal{Q}$ if L is (uniformly) $\breve{L}$-calibrated with respect to $\mathcal{Q}$.*

Fortunately, convex loss functions are always self-calibrated, as the following lemma shows.

Lemma 3.60 (Self-calibration of convex losses). *Every convex loss function $L : Y \times \mathbb{R} \to [0,\infty)$ is self-calibrated with respect to $\mathcal{Q}_{\min}(L)$.*

Proof. For a fixed distribution $\mathrm{Q} \in \mathcal{Q}_{\min}(L)$, we write $t_{\min} := \inf \mathcal{M}_{L,\mathrm{Q}}(0^+)$ and $t_{\max} := \sup \mathcal{M}_{L,\mathrm{Q}}(0^+)$. Now the map $t \mapsto \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}}$ is convex, and thus it is decreasing on $(-\infty,t_{\min}]$ and increasing on $[t_{\max},\infty)$. Furthermore, the convexity shows that $\mathcal{M}_{L,\mathrm{Q}}(0^+)$ is an interval and hence we find $\mathcal{M}_{\breve{L},\mathrm{Q}}(\varepsilon) = \{t \in \mathbb{R} : \breve{L}(\mathrm{Q},t) < \varepsilon\} = (t_{\min}-\varepsilon, t_{\max}+\varepsilon)$, $\varepsilon > 0$. This gives

$$\begin{aligned}
\delta_{\max,\breve{L},L}(\varepsilon,\mathrm{Q}) &= \inf_{t \notin \mathcal{M}_{\breve{L},\mathrm{Q}}(\varepsilon)} \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}} \\
&= \min\Bigl\{\mathcal{C}_{L,\mathrm{Q}}(t_{\min}-\varepsilon), \mathcal{C}_{L,\mathrm{Q}}(t_{\max}+\varepsilon)\Bigr\} - \mathcal{C}^*_{L,\mathrm{Q}} \qquad (3.68)\\
&> 0\,,
\end{aligned}$$

where we used the convention $\mathcal{C}_{L,\mathrm{Q}}(\pm\infty) := \infty$. □

It is easy to see by the results of Section 3.7 that in general convex losses are *not* uniformly self-calibrated. Therefore, we usually cannot expect strong inequalities in the sense of Theorem 3.22 for the self-calibration problem. However, the following theorem shows that for general self-calibrated losses, approximate risk minimizers approximate the Bayes decision functions in a weak sense. Its consequences for convex losses are discussed in Corollary 3.62.

Theorem 3.61 (Asymptotic self-calibration). *Let X be a complete measurable space, $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss that is self-calibrated with respect to some $\mathcal{Q} \subset \mathcal{Q}_{\min}(L)$, and P be a distribution of type $\mathcal{Q}$ with $\mathcal{R}^*_{L,\mathrm{P}} < \infty$. Then, for all $\varepsilon > 0$ and $\rho > 0$, there exists a $\delta > 0$ such that for all measurable $f : X \to \mathbb{R}$ satisfying $\mathcal{R}_{L,\mathrm{P}}(f) < \mathcal{R}^*_{L,\mathrm{P}} + \delta$ we have*

$$\mathrm{P}_X\Big(\big\{x \in X : \operatorname{dist}\big(f(x), \mathcal{M}_{L,\mathrm{P}(\,\cdot\,|x)}(0^+)\big) \geq \rho\big\}\Big) < \varepsilon\,.$$

Proof. For a fixed $\rho > 0$, we write $A_\rho = \{(\mathrm{Q},t) \in \mathcal{Q} \times \mathbb{R} : \breve{L}(\mathrm{Q},t) \geq \rho\}$. By Lemma 3.58, we then see that $\bar{L} := \mathbf{1}_{A_\rho}$ defines a template loss function whose P-instance $\bar{L}_\mathrm{P}$ is a detection loss with respect to $h := \mathbf{1}_X$ and $A := \{(x,t) \in X \times \mathbb{R} : \breve{L}(\mathrm{P}(\,\cdot\,|x),t) \geq \rho\}$. Furthermore, we have

$$\mathcal{M}_{\bar{L},\mathrm{Q}}(\varepsilon) = \big\{t \in \mathbb{R} : \breve{L}(\mathrm{Q},t) < \rho\big\} = \mathcal{M}_{\breve{L},\mathrm{Q}}(\rho)$$

for all $\varepsilon \in (0,1]$ and $\mathrm{Q} \in \mathcal{Q}$, and thus we obtain

$$\delta_{\max,\bar{L},L}(\varepsilon,\mathrm{Q}) = \delta_{\max,\breve{L},L}(\rho,\mathrm{Q}) \;>\; 0\,, \qquad\qquad \varepsilon \in (0,1],\, \mathrm{Q} \in \mathcal{Q}.$$

Since calibration functions are increasing, we then find that L is $\bar{L}_\mathrm{P}$-calibrated with respect to $\mathcal{Q}$. For $\varepsilon > 0$, Theorem 3.27 thus gives a $\delta > 0$ such that for $f : X \to \mathbb{R}$ with $\mathcal{R}_{L,\mathrm{P}}(f) < \mathcal{R}^*_{L,\mathrm{P}} + \delta$ we have

$$\mathrm{P}_X\big(\big\{x \in X : \breve{L}_\mathrm{P}(x, f(x)) \geq \rho\big\}\big) = \mathcal{R}_{\bar{L}_\mathrm{P},\mathrm{P}}(f) - \mathcal{R}^*_{\bar{L}_\mathrm{P},\mathrm{P}} < \varepsilon\,. \qquad \square$$

For convex losses L and distributions of $\mathcal{Q}_{1\text{-}\min}(L)$-type, we obtain the following consequence.

Corollary 3.62. *Let X be a complete measurable space, $L : Y \times \mathbb{R} \to [0,\infty)$ be a convex, supervised loss, and P be a distribution of type $\mathcal{Q}_{1\text{-}\min}(L)$ with $\mathcal{R}^*_{L,\mathrm{P}} < \infty$. Then there exists a P_X-almost surely unique minimizer $f^*_{L,\mathrm{P}}$ of $\mathcal{R}_{L,\mathrm{P}}$, and for all sequences (f_n) of measurable $f_n : X \to \mathbb{R}$, we have*

$$\mathcal{R}_{L,\mathrm{P}}(f_n) - \mathcal{R}^*_{L,\mathrm{P}} \to 0 \qquad \Longrightarrow \qquad f_n \to f^*_{L,\mathrm{P}} \quad \text{in probability } \mathrm{P}_X.$$

Proof. Lemma 3.12 together with the definition of $\mathcal{Q}_{1\text{-}\min}(L)$ shows that there exists a P_X-almost surely unique minimizer $f^*_{L,\mathrm{P}}$, and we thus find

$$\breve{L}_\mathrm{P}(x,t) = |t - f^*_{L,\mathrm{P}}(x)|\,, \qquad\qquad x \in X,\, t \in \mathbb{R}.$$

Theorem 3.61 together with Lemma 3.60 then yields $f_n \to f^*_{L,\mathrm{P}}$ in probability whenever the sequence (f_n) satisfies $\mathcal{R}_{L,\mathrm{P}}(f_n) \to \mathcal{R}^*_{L,\mathrm{P}}$. $\square$

Let us complete this discussion by describing situations in which we can replace the convergence in probability by a stronger notion of convergence.

Theorem 3.63 (Self-calibration inequalities). *Let X be a complete measurable space, $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss that is self-calibrated with respect to some $\mathcal{Q} \subset \mathcal{Q}_{1\text{-}\min}(L)$, and P be a distribution of type $\mathcal{Q}$ such that $\mathcal{R}^*_{L,\mathrm{P}} < \infty$. Moreover, assume that there exist a $p \in (0,\infty]$ and functions $b : X \to [0,\infty]$ and $\delta : [0,\infty) \to [0,\infty)$ such that*

$$\delta_{\max,\breve{L},L}(\varepsilon, \mathrm{P}(\,\cdot\,|x)) \geq b(x)\,\delta(\varepsilon)\,, \qquad \varepsilon > 0,\, x \in X,$$

and $b^{-1} \in L_p(\mathrm{P}_X)$. For a fixed $q \in (0,\infty)$, we define $\bar{\delta} : [0,\infty) \to [0,\infty)$ by

$$\bar{\delta}(\varepsilon) := \delta^{\frac{p}{p+1}}(\varepsilon^{1/q})\,, \qquad \varepsilon \in [0,\infty].$$

Then, for all measurable $f : X \to \mathbb{R}$ and $B_f := \|f - f^*_{L,\mathrm{P}}\|^q_\infty$, we have

$$\bar{\delta}^{**}_{B_f}\big(\|f - f^*_{L,\mathrm{P}}\|^q_{L_q(\mathrm{P}_X)}\big) \;\leq\; \|b^{-1}\|^{\frac{p}{p+1}}_{L_p(\mathrm{P}_X)}\big(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\big)^{\frac{p}{p+1}}\,,$$

where $\bar{\delta}^{**}_{B_f} : [0, B_f] \to [0,\infty]$ is the Fenchel-Legendre bi-conjugate of $\bar{\delta}_{|[0,B_f]}$.

Proof. We write $\hat{\delta}(\varepsilon) := \delta(\varepsilon^{1/q})$ for $\varepsilon \geq 0$, and $\bar{L} := \breve{L}^q$. Then $\bar{L}$ is a template loss by Lemma 3.58, and since $\mathcal{M}_{\breve{L},\mathrm{Q}}(\varepsilon) = \{t \in \mathbb{R} : \breve{L}(\mathrm{Q}, t) < \varepsilon\}$, we find

$$\delta_{\max,\bar{L},L}(\varepsilon, \mathrm{Q}) \;=\; \delta_{\max,\breve{L},L}(\varepsilon^{1/q}, \mathrm{Q}) \;\geq\; \hat{b}(x)\,\delta(\varepsilon) \qquad \varepsilon > 0,\, \mathrm{Q} \in \mathcal{Q}.$$

Moreover, we have $\hat{\delta}^{\frac{p}{p+1}} = \bar{\delta}$, and hence Theorem 3.25 applied to $\bar{L}$ and $\hat{\delta}$ yields the assertion. □

Note that if the function δ is of the form $\delta(\varepsilon) = \varepsilon^r$ for some $r > 0$ and we consider $q := \frac{pr}{p+1}$, then we obtain $\bar{\delta}(\varepsilon) = \varepsilon$. In this case, Theorem 3.63 yields

$$\|f - f^*_{L,\mathrm{P}}\|_{L_q(\mathrm{P}_X)} \;\leq\; \|b^{-1}\|^{1/r}_{L_p(\mathrm{P}_X)}\big(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\big)^{1/r}\,. \tag{3.69}$$

Moreover, if in this case we can only ensure $b^{-1} \in L_{p,\infty}(\mathrm{P}_X)$, then the norm $\|\cdot\|_{L_q(\mathrm{P}_X)}$ can be replaced by the Lorentz-norm $\|\cdot\|_{L_{q,\infty}(\mathrm{P}_X)}$ defined in Section A.5.5. For more details, we refer to Exercise 3.14.

The rest of this section applies the theory developed to some examples of practical importance. We begin with the problem of estimating the conditional probability $\mathrm{P}(y = 1|x)$ in classification, which has already been mentioned in the introduction of this section and which will be revisited in Section 8.5. To this end, we assume $Y := \{-1, 1\}$ in the following. Our first goal is to characterize situations when $\mathcal{Q}_{\min}(L) = \mathcal{Q}_Y$ for margin-based losses L.

Lemma 3.64 (Minimizers of margin-based losses). *Let L be a convex, margin-based loss represented by $\varphi : \mathbb{R} \to [0,\infty)$. Then we have $\mathcal{Q}_{\min}(L) = \mathcal{Q}_Y$ if and only if φ has a global minimum.*

Proof. If φ does not have a minimum, $\mathcal{C}_{L,1}(\,\cdot\,) = \varphi$ does not have a minimum, i.e., $\mathcal{M}_{L,1}(0^+) = \emptyset$. Conversely, if φ has a minimum, the same argument shows that $\mathcal{M}_{L,0}(0^+) = -\mathcal{M}_{L,1}(0^+) \neq \emptyset$. Therefore, let us fix an $\eta \in (0,1)$. If φ is constant, there is nothing to prove and hence we additionally assume that φ is not constant. The convexity of φ then shows that we have $\lim_{t\to\infty} \varphi(t) = \infty$ or $\lim_{t\to-\infty} \varphi(t) = \infty$. From this we immediately find $\mathcal{C}_{L,\eta}(t) \to \infty$ for $t \to \pm\infty$, and since $\mathcal{C}_{L,\eta}(\,\cdot\,)$ is continuous and convex, it thus has a global minimum. □

Together with Lemma 3.60, the preceding lemma immediately gives the following corollary that will be important when considering sparseness properties of support vector machines for classification in Section 8.5.

Corollary 3.65 (Self-calibration of margin-based losses). *Let L be a convex, margin-based loss whose representing function $\varphi : \mathbb{R} \to [0,\infty)$ has a global minimum. Then L is self-calibrated with respect to $\mathcal{Q}_Y$.*

With the help of Corollary 3.65, we see that the least squares loss and the (squared) hinge loss are self-calibrated with respect to $\mathcal{Q}_Y$, whereas the logistic loss is not. Furthermore, a simple calculation using Example 3.6 shows that the least squares loss is actually *uniformly* self-calibrated with respect to $\mathcal{Q}_Y$ and that the corresponding uniform self-calibration function is

$$\delta_{\max,\breve{L}_{\text{lsquares}},L_{\text{LS}}}(\varepsilon,\mathcal{Q}_Y) = \varepsilon^2\,, \qquad\qquad \varepsilon > 0\,.$$

However, neither the truncated least squares loss nor the hinge loss are uniformly self-calibrated with respect to $\mathcal{Q}_Y$, as we discuss in Exercise 3.15.

Let us now return to the problem of estimating the conditional probability $\eta(x) = \mathrm{P}(y = 1|x)$, $x \in X$. If we have a margin-based loss function L for which there is a one-to-one transformation between the sets of minimizers $\mathcal{M}_{L,\eta}(0^+)$ and η, then it seems natural to use self-calibration properties of L to investigate whether suitably transformed approximate L-risk minimizers approximate η. This approach is discussed in the following example.

Example 3.66. Exercise 3.2 shows that the **logistic loss for classification** $L_{\text{c-logist}}$ satisfies

$$\mathcal{M}_{L_{\text{c-logist}},\eta}(0^+) = \Big\{\ln\Big(\frac{\eta}{1-\eta}\Big)\Big\}\,, \qquad\qquad \eta \in (0,1).$$

In other words, if t_η^* denotes the element contained in $\mathcal{M}_{L_{\text{c-logist}},\eta}(0^+)$, then we have $\eta = \frac{1}{1+e^{-t_\eta^*}}$. Consequently, if t approximately minimizes $\mathcal{C}_{L_{\text{c-logist}},\eta}(\,\cdot\,)$, then it is close to t_η^* by Lemma 3.60 and hence $\frac{1}{1+e^{-t}}$ can serve as an estimate of η. However, investigating the quality of this estimate by the self-calibration function of $L_{\text{c-logist}}$ causes some technical problems since $L_{\text{c-logist}}$ is only self-calibrated with respect to the distributions $\mathrm{Q} \in \mathcal{Q}_Y$ with $\mathrm{Q}(\{1\}) \notin \{0,1\}$. Consequently, we now assess the quality of the estimate above *directly*. To this end, we introduce a new loss $L : \mathcal{Q}_Y \times \mathbb{R} \to [0,\infty)$, which we define by

$$L(\eta, t) := \Big|\eta - \frac{1}{1+e^{-t}}\Big|, \qquad \eta \in [0,1],\, t \in \mathbb{R}.$$

Then L is a template loss that measures the distance between η and its estimate $\frac{1}{1+e^{-t}}$. Let us compute the calibration function of $(L, L_{\text{c-logist}})$. To this end, we first observe that $\mathcal{C}^*_{L,\eta} = 0$ for all $\eta \in [0,1]$, and hence for $\varepsilon > 0$ an elementary calculation shows that

$$\begin{aligned}\mathcal{M}_{L,\eta}(\varepsilon) &= \{t \in \mathbb{R} : L(\eta, t) < \varepsilon\} \\ &= \left\{t \in \mathbb{R} : \ln \frac{(\eta-\varepsilon)_+}{1-\eta+\varepsilon} < t < \ln \frac{\eta+\varepsilon}{(1-\eta-\varepsilon)_+}\right\},\end{aligned}$$

where $(x)_+ := \max\{0, x\}$ for $x \in \mathbb{R}$ and $\ln 0 := -\infty$. For $\mathcal{C}_\eta(\infty) := \mathcal{C}_\eta(-\infty) := \infty$ and $\mathcal{C}_\eta(t) := \mathcal{C}_{L_{\text{c-logist}},\eta}(t) - \mathcal{C}^*_{L_{\text{c-logist}},\eta}$, Lemma 3.15 thus shows that

$$\delta_{\max,L,L_{\text{c-logist}}}(\varepsilon,\eta) = \min\left\{\mathcal{C}_\eta\left(-\ln\Big(\frac{1-\eta-\varepsilon}{\eta+\varepsilon}\Big)_+\right), \mathcal{C}_\eta\left(\ln\Big(\frac{\eta-\varepsilon}{1-\eta+\varepsilon}\Big)_+\right)\right\}.$$

From this we can conclude that $\delta_{\max,L,L_{\text{c-logist}}}(\varepsilon,\eta) = \delta_{\max,L,L_{\text{c-logist}}}(\varepsilon, 1-\eta)$ for all $\varepsilon \geq 0$, $\eta \in [0,1]$. Moreover, using the formulas of Exercise 3.2, we find

$$\mathcal{C}_\eta\left(\ln\Big(\frac{\eta-\varepsilon}{1-\eta+\varepsilon}\Big)_+\right) = \begin{cases}\eta \ln \frac{\eta}{\eta-\varepsilon} + (1-\eta)\ln\frac{1-\eta}{1-\eta+\varepsilon} & \text{if } \varepsilon < \eta \\ \infty & \text{otherwise}\end{cases}$$

and

$$\mathcal{C}_\eta\left(-\ln\Big(\frac{1-\eta-\varepsilon}{\eta+\varepsilon}\Big)_+\right) = \begin{cases}\eta \ln \frac{\eta}{\eta+\varepsilon} + (1-\eta)\ln\frac{1-\eta}{1-\eta-\varepsilon} & \text{if } \varepsilon < 1-\eta \\ \infty & \text{otherwise.}\end{cases}$$

In order to compare these expressions, let us write $g(\eta) := \eta \ln \frac{\eta}{\eta-\varepsilon} - \eta \ln \frac{\eta}{\eta+\varepsilon}$ for a fixed $\varepsilon \in (0, 1/2)$ and all η with $\varepsilon < \eta < 1-\varepsilon$. Then we have

$$g(1-\eta) = (1-\eta)\ln\frac{1-\eta}{1-\eta-\varepsilon} - (1-\eta)\ln\frac{1-\eta}{1-\eta+\varepsilon}$$

and

$$g'(\eta) = \frac{(\eta^2-\varepsilon^2)\ln\frac{\eta+\varepsilon}{\eta-\varepsilon} - 2\varepsilon\eta}{\eta^2-\varepsilon^2} =: \frac{g_\eta(\varepsilon)}{\eta^2-\varepsilon^2}\,.$$

Now observe that $g_\eta(0) = 0$ and $g'_\eta(\varepsilon) < 0$ for all $\varepsilon > 0$, and hence we obtain $g'(\eta) < 0$. Consequently, we have $g(\eta) \geq g(1-\eta)$, or in other words

$$\eta \ln \frac{\eta}{\eta-\varepsilon} + (1-\eta)\ln\frac{1-\eta}{1-\eta+\varepsilon} \;\geq\; \eta \ln \frac{\eta}{\eta+\varepsilon} + (1-\eta)\ln\frac{1-\eta}{1-\eta-\varepsilon}\,,$$

if and only if $\eta \leq \frac{1}{2}$. Therefore, for $\eta \in [0, 1/2]$, we find

$$\delta_{\max,L,L_{\text{c-logist}}}(\varepsilon,\eta) = \begin{cases}\eta \ln \frac{\eta}{\eta+\varepsilon} + (1-\eta)\ln\frac{1-\eta}{1-\eta-\varepsilon} & \text{if } \varepsilon < 1-\eta \\ \infty & \text{otherwise.}\end{cases}$$

In order to investigate whether $L_{\text{c-logist}}$ is L-calibrated with respect to $\mathcal{Q}_Y$, let us now find a simple lower bound of the calibration function above. To this end, let $h_\varepsilon(\eta) := \eta \ln \frac{\eta}{\eta+\varepsilon}$ for $\eta \in [0,1/2]$ and $\varepsilon \geq 0$. Then its derivative satisfies

$$h'_\varepsilon(\eta) = \ln \frac{\eta}{\eta+\varepsilon} + \frac{\varepsilon}{\eta+\varepsilon} = \ln\Bigl(1 - \frac{\varepsilon}{\eta+\varepsilon}\Bigr) + \frac{\varepsilon}{\eta+\varepsilon} \;\leq\; -\frac{\varepsilon}{\eta+\varepsilon} + \frac{\varepsilon}{\eta+\varepsilon} = 0\,,$$

and hence we find $\eta \ln \frac{\eta}{\eta+\varepsilon} \geq \frac{1}{2} \ln \frac{1}{1+2\varepsilon}$ for all $\eta \in [0,1/2]$, $\varepsilon \geq 0$. Analogously, we obtain $(1-\eta) \ln \frac{1-\eta}{1-\eta-\varepsilon} \geq \ln \frac{1}{1-\varepsilon}$ for $\eta \in [0,1/2]$, $\varepsilon \in [0,1-\eta)$. Both estimates together then yield

$$\delta_{\max,L,L_{\text{c-logist}}}(\varepsilon,\eta) \;\geq\; \frac{1}{2} \ln \frac{1}{1+2\varepsilon} + \ln \frac{1}{1-\varepsilon} \;\geq\; \varepsilon^2$$

for all $\eta \in [0,1/2]$ and all $\varepsilon \in [0,1-\eta)$. Consequently, $L_{\text{c-logist}}$ is *uniformly* L-calibrated with respect to $\mathcal{Q}_Y$, and the calibration function satisfies $\delta_{\max}(\varepsilon,\mathcal{Q}_Y) \geq \varepsilon^2$ for all $\varepsilon \geq 0$. For the loss function L^2, we thus obtain

$$\delta_{\max,L^2,L_{\text{c-logist}}}(\varepsilon,\mathcal{Q}_Y) = \delta_{\max,L,L_{\text{c-logist}}}(\varepsilon^{1/2},\mathcal{Q}_Y) \geq \varepsilon\,, \qquad\qquad \varepsilon \geq 0.$$

By Theorem 3.22, we then see that for all measurable $f : X \to \mathbb{R}$ we have

$$\int_X \Bigl|\eta(x) - \frac{1}{1+e^{-f(x)}}\Bigr|^2 \, d\mathrm{P}_X(x) \;\leq\; \mathcal{R}_{L_{\text{c-logist}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{c-logist}},\mathrm{P}}\,,$$

i.e., we can assess the quality of the estimate $\frac{1}{1+e^{-f(x)}}$ in terms of $\|\cdot\|_2$. ◁

Our last goal is to investigate the self-calibration properties of the τ-pinball loss $L_{\tau\text{-pin}}$. Proposition 3.9 showed that the minimizer of this convex supervised loss was the τ-quantile, and consequently $L_{\tau\text{-pin}}$ can be used to estimate the conditional τ-quantile. However, so far we only have a rather weak justification in the sense of Theorem 3.61. The following example discusses some conditions on the distribution P, which provides a stronger justification.

Example 3.67. For fixed $\tau \in (0,1)$, let $L := L_{\tau\text{-pin}}$ be the **τ-pinball loss** defined in Example 2.43. Furthermore, let Q be a distribution on $\mathbb{R}$ such that $|\mathrm{Q}|_1 < \infty$ and let t^* be a τ-quantile of Q, i.e., we simultaneously have

$$\mathrm{Q}\bigl((-\infty,t^*]\bigr) \geq \tau \qquad\qquad \text{and} \qquad\qquad \mathrm{Q}\bigl([t^*,\infty)\bigr) \geq 1-\tau\,. \tag{3.70}$$

If t^* is the only τ-quantile of Q, i.e., t^* is uniquely defined by (3.70), then the formulas of Proposition 3.9 show

$$\delta_{\max,\check{L},L}(\varepsilon,\mathrm{Q}) = \min\biggl\{\varepsilon q_+ + \int_0^\varepsilon \mathrm{Q}\bigl((t^*,t^*+s)\bigr)\,ds,\, \varepsilon q_- + \int_0^\varepsilon \mathrm{Q}\bigl((t^*-s,t^*)\bigr)\,ds\biggr\}$$

for all $\varepsilon \geq 0$, where q_+ and q_- are the real numbers found in Proposition 3.9.

Let us now denote the set of all distributions Q for which the inequalities in (3.70) strictly hold by $\mathcal{Q}_\tau^{>0}$. For $\mathrm{Q} \in \mathcal{Q}_\tau^{>0}$, we then have $\min\{q_+, q_-\} > 0$ and hence t^* is uniquely determined. Moreover, the self-calibration function satisfies

$$\delta_{\max,\check{L},L}(\varepsilon, \mathrm{Q}) \geq c_\mathrm{Q}\, \varepsilon\,, \qquad \varepsilon \geq 0, \tag{3.71}$$

where $c_\mathrm{Q} := \min\{q_+, q_-\}$. For a fixed distribution P of $\mathcal{Q}_\tau^{>0}$-type, we now define the function $b : X \to [0, \infty)$ by $b(x) := c_{\mathrm{P}(\,\cdot\,|x)}$, $x \in X$, where $c_{\mathrm{P}(\,\cdot\,|x)}$ denotes the constant in (3.71), which belongs to the conditional distribution $\mathrm{P}(\,\cdot\,|x)$. If we have $b^{-1} \in L_p(\mathrm{P}_X)$, then Theorem 3.63, see also (3.69), shows

$$\|f - f^*_{\tau,\mathrm{P}}\|_{L_q(\mathrm{P}_X)} \leq \|b^{-1}\|_{L_p(\mathrm{P}_X)} \left(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\right) \tag{3.72}$$

for all measurable functions $f : X \to \mathbb{R}$, where $f^*_{\tau,\mathrm{P}}(x)$ denotes the τ-quantile of $\mathrm{P}(\,\cdot\,|x)$ and $q := \frac{p}{p+1}$.

Although (3.72) provides a nice relationship between the excess pinball risk and our goal of estimating the conditional quantile function $f^*_{\tau,\mathrm{P}}$, the distributions P of $\mathcal{Q}_\tau^{>0}$-type seem a bit unrealistic for practical situations. Therefore, let us finally consider a more realistic scenario. To this end, we fix an $\alpha > 0$ and say that a distribution Q with $|\mathrm{Q}|_1 < \infty$ is of type $\mathcal{Q}_\tau^\alpha$ if there exists a τ-quantile t^* of Q and a constant $c_\mathrm{Q} > 0$ such that

$$\mathrm{Q}\big((t^*, t^* + s)\big) \geq c_\mathrm{Q}\, s \qquad \text{and} \qquad \mathrm{Q}\big((t^* - s, t^*)\big) \geq c_\mathrm{Q}\, s \tag{3.73}$$

for all $s \in [0, \alpha]$. Obviously, for such distributions, the τ-quantile t^* is uniquely determined. Moreover, if Q has a density h_Q with respect to the Lebesgue measure and this density satisfies $h_\mathrm{Q}(t) \geq c_\mathrm{Q}$ for all $t \in [t^* - \alpha, t^* + \alpha]$, then Q is of type $\mathcal{Q}_\tau^\alpha$. Let us now define $\delta : [0, \infty) \to [0, \infty)$ by

$$\delta(\varepsilon) := \begin{cases} \varepsilon^2/2 & \text{if } \varepsilon \in [0, \alpha] \\ \alpha\varepsilon - \alpha^2/2 & \text{if } \varepsilon > \alpha\,. \end{cases}$$

Then a simple calculation yields

$$\delta_{\max,\check{L},L}(\varepsilon, \mathrm{Q}) \geq c_\mathrm{Q}\delta(\varepsilon)\,, \qquad \varepsilon \geq 0,$$

for all $\mathrm{Q} \in \mathcal{Q}_\tau^\alpha$, where c_Q is the constant satisfying (3.73). For fixed $p \in (0, \infty]$, we further define $\bar{\delta} : [0, \infty) \to [0, \infty)$ by $\bar{\delta}(\varepsilon) := \delta^{\frac{p}{p+1}}(\varepsilon^{\frac{p+1}{p}})$, $\varepsilon \geq 0$. In view of Theorem 3.63, we then need to find a convex function $\hat{\delta} : [0, \infty) \to [0, \infty)$ such that $\hat{\delta} \leq \bar{\delta}$. To this end, we define

$$\hat{\delta}(\varepsilon) := \begin{cases} s_p^p \varepsilon^2 & \text{if } \varepsilon \in \big[0, s_p a_p\big] \\ a_p\big(\varepsilon - s_p^{p+2} a_p\big) & \text{if } \varepsilon > s_p a_p\,, \end{cases}$$

where $a_p := \alpha^{p/(p+1)}$ and $s_p := 2^{-1/(p+1)}$. An easy calculation shows that $\hat{\delta} : [0, \infty) \to [0, \infty)$ is continuously differentiable with non-decreasing derivative.

Consequently, $\hat{\delta}$ is convex. Moreover, since $(\varepsilon^{\frac{p+1}{p}} - \alpha/2)^{-\frac{1}{p+1}} \varepsilon^{\frac{1}{p}} \geq 1$ we have $\hat{\delta}' \leq \bar{\delta}'$ and hence we find $\hat{\delta} \leq \bar{\delta}$ by the fundamental theorem of calculus.

For a distribution P of $\mathcal{Q}_\tau^\alpha$-type, we now define the function $b : X \to [0, \infty)$ by $b(x) := c_{\mathrm{P}(\cdot|x)}$, $x \in X$, where $c_{\mathrm{P}(\cdot|x)}$ is determined by (3.73). If b satisfies $b^{-1} \in L_p(\mathrm{P}_X)$ for some $p \in (0, \infty]$, Theorem 3.63 together with our considerations above shows

$$\|f - f^*_{\tau,\mathrm{P}}\|_{L_q(\mathrm{P}_X)} \leq \sqrt{2}\, \|b^{-1}\|^{1/2}_{L_p(\mathrm{P}_X)} \left(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\right)^{1/2} \tag{3.74}$$

for $q := \frac{p}{p+1}$ and all $f : X \to \mathbb{R}$ satisfying $\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \leq 2^{-\frac{p+2}{p+1}} \alpha^{\frac{2p}{p+1}}$. $\triangleleft$

3.10 Further Reading and Advanced Topics

The idea of using a surrogate loss developed quite independently in statistics and machine learning. Indeed, in statistics, its development was mainly motivated by the search for more robust estimation methods (see, e.g., Huber, 1964), in particular for regression problems. On the other hand, in machine learning, surrogate losses were mainly considered as a trick to find faster classification algorithms. However, only very recently has the relation between the risks of these surrogates and the classification risk been investigated. The first observations on the set of minimizers were made by Lin (2002b). Later he (see Lin, 2004, Theorem 3.1 and Lemma 4.1) established a result somewhat similar to Theorem 3.36 and a bound on the excess classification risk that generalizes the widely known Theorem 2.2 from Devroye *et al.* (1996). Independently of Lin, Zhang (2004b) established the first general inequalities between the excess classification risk and the excess risks of margin-based surrogate losses. Furthermore, he mentioned that some applications also require estimating the conditional probability and concludes that some margin-based losses, including the hinge loss, are not suited for this task. Another independent result, established by Steinwart (2005), gives a sufficient condition for continuous, supervised losses L that ensures an asymptotic relation (in the sense of Question 3.1) between the excess classification risk and the excess L-risk. However, the big breakthrough in understanding surrogate margin-based losses was then made by Bartlett *et al.* (2006). In fact, all the main results on classification calibrated, margin-based losses presented in Section 3.4 were shown by these authors, though condition (3.40) was already investigated by Mammen and Tsybakov (1999), and Tsybakov (2004) in the context of density level detection. We refer to Steinwart *et al.* (2005) and Steinwart (2007), who translated their findings into the language of calibration inequalities.

Prior to Steinwart (2007), the only result for weighted classification (also known as *cost-sensitive classification*) that deals with calibration issues was presented by Lin *et al.* (2002), though weighted classification itself had been considered earlier by, e.g., Elkan (2001). The presentation in Section 3.5 closely

follows Steinwart's work. Furthermore, there are recent results on surrogates for multi-class classification that we have not presented here due to lack of space. For more information, we refer to Lee *et al.* (2004), Zhang (2004a), Tewari and Bartlett (2005), and the references therein.

Proposition 3.44, which shows the unique role of the least squares loss for estimating the regression function, was independently found by Caponnetto (2005) and Steinwart (2007). Besides the basic notions and examples, the rest of Section 3.7 is based on the work of Steinwart (2007). Finally, it is worth mentioning that the approach in Section 3.7 substantially differs from the traditional maximum-likelihood motivation for the least squares loss already used by Gauss. We refer to Schölkopf and Smola (2002) for a brief introduction to the maximum-likelihood motivation and to Kardaun *et al.* (2003) for a discussion on this motivation.

The asymptotic theory on surrogate losses developed in Section 3.2 is a generalization of the results of Steinwart (2005). Moreover, the inequalities for general surrogate losses established in Section 3.3 were deeply inspired by the work of Bartlett *et al.* (2006). However, the key results of this section, namely Theorem 3.22 and Theorem 3.25, can also be derived from Theorem 24 of Zhang (2004a). Finally, a self-calibration result for classification calibrated surrogates similar to Theorem 3.61 was already shown by Steinwart (2003). In the presentation of all of these results, we closely followed Steinwart (2007).

3.11 Summary

In this chapter, we developed a general theory that allows us to *a)* identify suitable surrogate loss functions and *b)* relate the excess risks of such surrogate losses with the excess risks of the original (target) loss function. The main concept of this theory was the *calibration function*, which compares the *inner* excess risks of the losses involved. With the help of the calibration function, we then introduced the notions of calibration and uniform calibration, which (essentially) characterize how the excess risks involved can be compared. We then applied the general theory to some important learning scenarios:

- **Classification.** Here we showed that, for margin-based losses, calibration and uniform calibration are equivalent concepts. Furthermore, we developed a way to establish inequalities between the excess classification risk and the excess risk of margin-based losses. We then established an easy test to check whether a given *convex*, margin-based loss function is classification calibrated. Finally, we further simplified the computation of the uniform calibration function for such losses.
- **Weighted classification.** We showed that a simple weighting method for classification calibrated, margin-based loss functions produces loss functions that are calibrated to the weighted classification scenario. With the help of this weighting method, we then translated the major results on

unweighted classification calibration into analogous results on weighted classification calibration.

- **Regression.** Here we first showed that the least squares loss is essentially the only distance-based loss that can be used to find the regression function if one only knows that the average second moment of the noise distributions is finite. For some large classes of *symmetric* noise distributions, we then characterized (uniformly) least squares calibrated losses. Here it turned out that the convexity and related stronger notions play a crucial role. In particular, we showed that for symmetric, unbounded noise every uniformly least squares calibrated and symmetric loss must grow at least as fast as the least squares loss, and consequently one cannot avoid assuming the finiteness of the second moments for such distributions and losses. Furthermore, we have seen that for slower-growing losses, such as the absolute distance loss, the latter requirement can be replaced by non-parametric assumptions on the concentration around the mean.
- **Density level detection.** We first showed that the DLD learning scenario can be treated as a supervised learning problem that is similar to a classification problem. It then turned out that every classification calibrated loss is DLD-calibrated. However, unlike for classification, there exists *no* uniformly DLD-calibrated supervised loss, and consequently it is impossible to establish inequalities between the DLD-risk and excess risks of supervised surrogates without further assumptions on the density.
- **Self-calibration.** It is of both theoretical and practical interest whether approximate risk minimizers approximate the true risk minimizer. In Section 3.9, we developed a general framework to investigate this issue. In particular, we showed that convex losses always guarantee a weak positive result. Finally, we applied the general theory to the logistic loss for classification and the pinball loss.

The theory developed and its consequences for the learning scenarios above will play an important role when we investigate the corresponding kernel-based learning procedures in later chapters. However, it is worth mentioning that the results of this chapter are *algorithm independent*, i.e., they can be used for any algorithm whose *surrogate* risk performance is understood.

3.12 Exercises

3.1. Inner risks of the squared hinge loss ($\star$)
Recall that in Example 2.28 we defined the squared hinge loss by $L(y,t) := (\max\{0, 1-yt\})^2$, $y = \pm 1$, $t \in \mathbb{R}$. Using the definitions in (3.8), show that for $\eta \in [0,1]$ we have $\mathcal{C}^*_{L,\eta} = 4\eta(1-\eta)$ and

$$\mathcal{M}_{L,\eta}(0^+) = \begin{cases} (-\infty, -1] & \text{if } \eta = 0 \\ \{2\eta - 1\} & \text{if } 0 < \eta < 1 \\ [1, \infty) & \text{if } \eta = 1\,. \end{cases}$$

Moreover, show that, for $\eta \in [1/2, 1]$ and $t \in \mathbb{R}$, the excess inner risk can be computed by

$$\mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} = \begin{cases} 4\eta^2 - 3\eta - 2\eta t + \eta t^2 & \text{if } t \le -1 \\ (t - 2\eta + 1)^2 & \text{if } t \in [-1, 1] \\ (1-\eta)(1 + 2t + t^2 - 4\eta) & \text{if } t \ge 1\,. \end{cases}$$

3.2. Logistic loss for classification($\star$)
Recall that in Example 2.29 we defined the logistic loss for classification by $L(y,t) := \ln(1+\exp(-yt))$, $y = \pm 1$, $t \in \mathbb{R}$. Show the following formulas using the notations in (3.8) and the convention $0 \ln 0 := 0$:

$$\begin{aligned} \mathcal{C}^*_{L,\eta} &= -\eta \ln(\eta) - (1-\eta)\ln(1-\eta)\,, \\ \mathcal{M}_{L,\eta}(0^+) &= \big\{\ln(\eta) - \ln(1-\eta)\big\}\,, \qquad \text{if } \eta \ne 0, 1, \\ \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} &= \eta \ln\big(\eta(1+e^{-t})\big) + (1-\eta)\ln\big((1-\eta)(1+e^t)\big). \end{aligned}$$

3.3. Calibration function ($\star$)
Let $L_{\text{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\text{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be loss functions, Q be a distribution on Y, and $x \in X$ with $\mathcal{C}^*_{L_{\text{tar}},\mathrm{Q},x} < \infty$ and $\mathcal{C}^*_{L_{\text{sur}},\mathrm{Q},x} < \infty$. Assume that $\delta : [0,\infty] \to [0,\infty]$ is an increasing function with

$$\delta\big(\mathcal{C}_{L_{\text{tar}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\text{tar}},\mathrm{Q},x}\big) \;\le\; \mathcal{C}_{L_{\text{sur}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\text{sur}},\mathrm{Q},x}\,, \qquad t \in \mathbb{R}.$$

Show that $\delta(\varepsilon) \le \delta_{\max}(\varepsilon, \mathrm{Q}, x)$ for all $\varepsilon \in [0,\infty]$.

Hint: Assume the converse and use Lemma 3.14.

3.4. Characterization of calibration ($\star\star\star$)
Prove Corollary 3.19.

Hint for ii) $\Rightarrow$ *i):* Assume that L_{sur} is not L_{tar}-calibrated to construct a "simple" distribution P that violates *ii*). Furthermore, use that the condition $\mathcal{R}^*_{L_{\text{tar}},\mathrm{P}} < \infty$ is automatically satisfied since L_{tar} is bounded.

3.5. Uniform calibration function ($\star\star$)
Let $L_{\text{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\text{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be loss functions and $\mathcal{Q}$ be a set of distributions on Y. Show that for all $\varepsilon \in [0,\infty]$ we have

$$\delta_{\max}(\varepsilon, \mathcal{Q}) = \max\big\{\delta \ge 0 : \mathcal{M}_{L_{\text{sur}},\mathrm{Q},x}(\delta) \subset \mathcal{M}_{L_{\text{tar}},\mathrm{Q},x}(\varepsilon) \text{ for all } \mathrm{Q} \in \mathcal{Q},\, x \in X\big\}\,.$$

3.6. Uniformly calibrated supervised losses ($\star\star\star\star$)
Let $L_{\text{tar}} : Y \times \mathbb{R} \to [0,\infty)$ and $L_{\text{sur}} : Y \times \mathbb{R} \to [0,\infty)$ be supervised loss functions, X be a complete measurable space, and μ be a probability measure on X. Assume that there exist mutually disjoint measurable subsets $A_n \subset X$ with $\mu(A_n) > 0$ for all $n \in \mathbb{N}$. Finally, let $\mathcal{Q}$ be a set of distributions on Y such that $\mathcal{C}^*_{L_{\text{tar}},\mathrm{Q}} < \infty$ and $\mathcal{C}^*_{L_{\text{sur}},\mathrm{Q}} < \infty$ for all $\mathrm{Q} \in \mathcal{Q}$. Show that there exists a distribution P on $X \times Y$ of type $\mathcal{Q}$ such that $\mathrm{P}_X = \mu$ and

$$\delta_{\max}(\varepsilon, \mathcal{Q}) = \inf_{x \in X} \delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x)\,, \qquad \varepsilon \in [0, \infty]. \tag{3.75}$$

Hint: First show that $U := \{\varepsilon > 0 : \delta_{\max}(\,\cdot\,, \mathcal{Q}) \text{ not continuous at } \varepsilon\}$ is at most countable. Then show equation (3.75) for an enumeration $(\varepsilon_n)_{n \in \mathbb{N}}$ of $U \cup \{r \in \mathrm{Q} : r \geq 0\}$. Use this to conclude the general case.

3.7. Characterization of calibration for detection losses ($\star\star\star$)
Prove Theorem 3.27 using the same idea as in Exercise 3.4.

3.8. Some more margin-based losses ($\star\star$)
Determine the calibration function with respect to the classification loss for the *exponential loss* given by $\varphi(t) := \exp(-t)$, $t \in \mathbb{R}$, and the *sigmoid loss* given by $\varphi(t) := 1 - \tanh t$, $t \in \mathbb{R}$. Is the latter classification calibrated?

3.9. Inequalities for unweighted classification ($\star\star$)
Use Theorems 3.34 and 3.22 to establish inequalities between the excess classification risk and the excess L-risk for L being the least squares loss, the hinge loss, the squared hinge loss, and the logistic loss for classification. How do these inequalities change when we additionally assume (3.40)?

3.10. Another weighted classification scenario ($\star\star\star$)
Let $h : X \to [0, \infty)$ be measurable. For the loss $L : X \times Y \times \mathbb{R} \to [0, \infty)$ defined by $L(x, y, t) := h(x) L_{\text{class}}(y, t)$, perform the following tasks:

i) Investigate which margin-based losses are L-calibrated.
ii) When are L-calibrated margin-based losses uniformly L-calibrated?
iii) Given a margin-based loss represented by some φ, determine the calibration function for the loss $(x, y, t) \mapsto h(x)\varphi(yt)$. Compare the results with those for the unweighted version.
iv) Find some practical situations in which L may be of interest.

3.11. Asymptotic relation between excess risks revisited ($\star\star$)
Show that in general a strictly positive calibration function is *not* sufficient for the implication (3.18).

Hint: Assume that L_{LS} is the target loss and that $L_{p\text{-dist}}$ is the surrogate loss for some $p \in [1, 2)$. Furthermore, consider the distribution P on $[0, 1) \times \mathbb{R}$ with P_X being the uniform distribution and $\mathrm{P}(\,\cdot\,|x) = \delta_{\{0\}}$ for all $x \in [0, 1]$.

3.12. Modulus of convexity for p-th power distance loss ($\star\star\star$)
For $p \in (1, 2)$, define $\psi : \mathbb{R} \to [0, \infty)$ by $\psi(t) := |t|^p$, $t \in \mathbb{R}$. Show for all $B > 0$ and $\varepsilon \in [0, B]$ that

$$\frac{p(p-1)}{2} B^{p-2} \varepsilon^2 \;\leq\; \delta_{\psi|_{[-B,B]}}(2\varepsilon) \;\leq\; \frac{p}{2(p-1)^2} B^{p-2} \varepsilon^2\,.$$

Hint: First show a $s^{a-1} \leq s^a - (s-1)^a \leq s^{a-1}$ for all $0 < a < 1$ and all $s \geq 1$. Use this to estimate $\psi'(t) - \psi'(t - \varepsilon)$, and then apply Lemma A.6.19.

3.13. Reverse calibration for DLD ($\star\star\star$)
Let μ be a probability measure on a measurable space X and $Y := \{-1, 1\}$. Furthermore, let $\rho > 0$ and $g : X \to [0, \infty)$ be a measurable function with $\|g\|_{L_1(\mu)} = 1$. Then, for $\mathrm{P} := g\mu \ominus_\rho \mu$ and all sequences (f_n) of measurable functions $f_n : X \to \mathbb{R}$, we have

$$\mathcal{R}_{L_{\mathrm{DLD}},\mathrm{P}}(f_n) \to 0 \qquad \Longrightarrow \qquad \mathcal{R}_{L_{\mathrm{class}},\mathrm{P}}(f_n) \to \mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P}}\,.$$

Hint: Compute the calibration function $\delta_{\max,L_{\mathrm{class}},\bar{L}_{\mathrm{DLD}}}(\,\cdot\,,\,\cdot\,)$ using Lemma 3.32. Then observe that $\mu(\{x \in X : \eta(x) = 1\}) = 0$ and use Corollary 3.19.

3.14. Another inequality for self-calibrated losses ($\star\star\star$)
Let $L : Y \times \mathbb{R} \to [0, \infty)$ be a supervised loss that is self-calibrated with respect to some $\mathcal{Q} \subset \mathcal{Q}_{\min}(L)$ and P be a distribution on $X \times Y$ that is of type $\mathcal{Q}$. Assume further that there exist $p > 0$, $q > 0$, and a function $b : X \to [0, \infty]$ with $b^{-1} \in L_{p,\infty}(\mathrm{P}_X)$ and

$$\delta_{\max,\check{L}_{\mathrm{P}},L}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) \geq \varepsilon^q\, b(x)\,, \qquad \varepsilon > 0,\ x \in X.$$

Show that for all measurable $f : X \to \mathbb{R}$ we have

$$\mathrm{P}_X\big(\big\{x \in X : \check{L}_{\mathrm{P}}\big(x, f(x)\big) \geq \rho\big\}\big) \;\leq\; 2\left(\frac{\|b^{-1}\|_p\,\big(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\big)}{\rho^q}\right)^{\frac{p}{p+1}}.$$

If in addition $\mathcal{R}_{L,\mathrm{P}}(\,\cdot\,)$ has an almost surely unique minimizer $f^*_{L,\mathrm{P}}$, interpret the result in terms of Lorentz norms and compare it with Theorem 3.63.

Hint: Use the set A_ρ from the proof of Theorem 3.61 and apply Theorem 3.28.

3.15. Self-calibration of the (squared) hinge loss ($\star\star\star$)
i) Show that the self-calibration function of the hinge loss is given by

$$\delta_{\max,\check{L}_{\mathrm{hinge}},L_{\mathrm{hinge}}}(\varepsilon, \eta) = \begin{cases} \varepsilon\,\min\{\eta, 1-\eta, 2\eta - 1\} & \text{if } \eta \neq 0, 1/2, 1 \\ \varepsilon & \text{if } \eta \in \{0, 1\} \\ \infty & \text{if } \eta = 1/2 \end{cases}$$

for all $\varepsilon \in (0, 2]$, $\eta \in [0, 1]$. Is the hinge loss uniformly self-calibrated?
ii) Show that, for all distributions P on $X \times Y$ and all $p \in (0, \infty)$ and $\varepsilon > 0$, there exists a $\delta > 0$ such that for all measurable $f : X \to \mathbb{R}$ we have

$$\mathcal{R}_{L_{\mathrm{hinge}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{hinge}},\mathrm{P}} \;\leq\; \delta \quad \Longrightarrow \quad \|x \mapsto \check{L}_{\mathrm{hinge},\mathrm{P}}(x, \widehat{f}(x))\|_{L_p(\mathrm{P}_X)} \;\leq\; \varepsilon,$$

where the clipping is at ± 1. Compare this with Theorem 3.61. Find conditions on P such that Theorem 3.25 gives inequalities for clipped functions.
iii) Use Exercise 3.1 and Equation (3.68) to show that the squared hinge loss is *not* uniformly self-calibrated.

Hint: For the first implication in *ii)* use Theorem 3.17.