2

Loss Functions and Their Risks

Overview. *We saw in the introduction that the learning problems we consider in this book can be described by loss functions and their associated risks. In this chapter, we present some common examples of such learning problems and introduce a general notion for losses and their risks. Furthermore, we discuss some elementary yet fundamental properties of these concepts.*

Prerequisites. *Basic knowledge of measure and integration theory provided in Section A.3.*

Usage. *Sections 2.1 and 2.2 are essential for the rest of this book, and Sections 2.3 and 2.4 are used whenever we deal with classification and regression problems, respectively.*

Every learning problem requires that we specify our learning goal, i.e., what we ideally would like to achieve. We saw in the introduction that the specification of the learning problems treated in this book needs a *loss* $L(x, y, f(x))$ that describes the cost of the discrepancy between the prediction $f(x)$ and the observation y at the point x. To the loss L we then associate a *risk* that is defined by the average future loss of f. This chapter introduces these concepts and presents important examples of learning goals described by losses. In addition, basic yet useful properties of risks are derived from properties of the corresponding losses.

2.1 Loss Functions: Definition and Examples

In this section, we will first introduce loss functions and their associated risks. We will then present some basic examples of loss functions that describe the most important learning scenarios we are dealing with in this book.

In order to avoid notational overload, we assume throughout this chapter that subsets of $\mathbb{R}^d$ are equipped with their Borel σ-algebra and that products of measurable spaces are equipped with the corresponding product σ-algebra.

Let us now recall from the introduction that we wish to find a function $f : X \to \mathbb{R}$ such that for $(x, y) \in X \times Y$ the value $f(x)$ is a good prediction of y at x. The following definition will help us to define what we mean by "good".

Definition 2.1. *Let $(X, \mathcal{A})$ be a measurable space and $Y \subset \mathbb{R}$ be a closed subset. Then a function $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is called a* ***loss function****, or simply a* ***loss****, if it is measurable.*

In the following, we will interpret $L(x, y, f(x))$ as the *cost*, or *loss*, of predicting y by $f(x)$ if x is observed, i.e., the smaller the value $L(x, y, f(x))$ is, the better $f(x)$ predicts y in the sense of L. From this it becomes clear that constant loss functions, such as $L := 0$, are rather meaningless for our purposes, since they do not distinguish between good and bad predictions.

Let us now recall from the introduction that our major goal is to have a small *average* loss for future unseen observations (x, y). This leads to the following definition.

Definition 2.2. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss function and P be a probability measure on $X \times Y$. Then, for a measurable function $f : X \to \mathbb{R}$, the* ***L-risk*** *is defined by*

$$\mathcal{R}_{L,\mathrm{P}}(f) := \int_{X \times Y} L\big(x, y, f(x)\big)\, d\mathrm{P}(x, y) = \int_X \int_Y L\big(x, y, f(x)\big)\, d\mathrm{P}(y|x)\, d\mathrm{P}_X(x).$$

Note that the function $(x, y) \mapsto L(x, y, f(x))$ is measurable by our assumptions, and since it is also non-negative, the above integral over $X \times Y$ always exists, although it is not necessarily finite. In addition, our label space $Y \subset \mathbb{R}$ is closed, and hence Lemma A.3.16 ensures the existence of the *regular* conditional probability $\mathrm{P}(\,\cdot\,|x)$, used in the inner integral.

For a given sequence $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$, we write $\mathrm{D} := \frac{1}{n} \sum_{i=1}^n \delta_{(x_i, y_i)}$, where $\delta_{(x_i, y_i)}$ denotes the Dirac measure at (x_i, y_i). In other words, D is the empirical measure associated to D. The risk of a function $f : X \to \mathbb{R}$ with respect to this measure is called the **empirical L-risk**

$$\mathcal{R}_{L,\mathrm{D}}(f) = \frac{1}{n} \sum_{i=1}^n L\big(x_i, y_i, f(x_i)\big)\,. \tag{2.1}$$

Let us now assume for a moment that D is a sequence of i.i.d. observations generated by P and f satisfies $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$. Recalling the law of large numbers, we then see that the empirical risk $\mathcal{R}_{L,\mathrm{D}}(f)$ is close to $\mathcal{R}_{L,\mathrm{P}}(f)$ with high probability. In this sense, the L-risk of f can be seen as an approximation to the average loss on the observations D (and vice versa).

Now recall that $L(x, y, f(x))$ was interpreted as a cost that we wish to keep small, and hence it is natural to look for functions f whose risks are as small as possible. Since the smallest possible risk plays an important role throughout this book, we now formally introduce it.

Definition 2.3. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss function and P be a probability measure on $X \times Y$. Then the minimal L-risk*

$$\mathcal{R}^*_{L,\mathrm{P}} := \inf\big\{\mathcal{R}_{L,\mathrm{P}}(f) \,\big|\, f : X \to \mathbb{R} \text{ measurable}\big\}$$

is called the ***Bayes risk*** *with respect to* P *and* L*. In addition, a measurable* $f^*_{L,\mathrm{P}} : X \to \mathbb{R}$ *with* $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) = \mathcal{R}^*_{L,\mathrm{P}}$ *is called a* ***Bayes decision function****.*

Usually the first step in solving a practical learning problem is finding a loss function that best describes the often only informally specified learning goal. In general, the choice of a suitable loss function strongly depends on the specific application, and hence only a few general statements are possible in this regard. However, there are a few basic learning scenarios that often fit the learning problem at hand, and hence we will formally introduce these scenarios and their corresponding loss functions now.

Example 2.4 (Standard binary classification). Let $Y := \{-1, 1\}$ and P be an unknown data-generating distribution on $X \times Y$. Then the informal goal in (binary) classification is to predict the label y of a pair (x, y) drawn from P if only x is observed. The most common loss function describing this learning goal is the **classification loss**[1] $L_{\text{class}} : Y \times \mathbb{R} \to [0, \infty)$, which is defined by

$$L_{\text{class}}(y, t) := \mathbf{1}_{(-\infty,0]}\big(y \operatorname{sign} t\big)\,, \qquad y \in Y,\, t \in \mathbb{R}\,, \tag{2.2}$$

where we use the convention $\operatorname{sign} 0 := 1$. Note that L_{class} only penalizes predictions t whose signs disagree with that of y, so it indeed reflects our informal learning goal. Now, for a measurable function $f : X \to \mathbb{R}$, an elementary calculation shows

$$\begin{aligned}\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f) &= \int_X \eta(x)\mathbf{1}_{(-\infty,0)}(f(x)) + (1 - \eta(x))\mathbf{1}_{[0,\infty)}(f(x))\, d\mathrm{P}_X(x) \\ &= \mathrm{P}\big(\{(x, y) \in X \times Y : \operatorname{sign} f(x) \neq y\}\big)\,,\end{aligned}$$

where $\eta(x) := \mathrm{P}(y = 1|x)$, $x \in X$. From this we conclude that f is a Bayes decision function if and only if $(2\eta(x) - 1) \operatorname{sign} f(x) \geq 0$ for P_X-almost all $x \in X$. In addition, this consideration yields

$$\mathcal{R}^*_{L_{\text{class}},\mathrm{P}} = \int_X \min\{\eta, 1 - \eta\}\, d\mathrm{P}_X\,. \qquad \lhd$$

The loss function L_{class} equally weights both types of errors, namely $y = 1$ while $f(x) < 0$, and $y = -1$ while $f(x) \geq 0$. This particularly makes sense in situations in which one wishes to *categorize* objects such as hand-written characters or images. In many practical situations, however, both error types should be weighted differently. For example, if one wants to detect computer network intrusions, then depending on the available resources for investigating alarms and the sensitivity of the network, the two types of errors, namely false alarms and undetected intrusions, are likely to have different actual costs.

[1] Formally, L_{class} is not a loss function; however, we can canonically identify it with the loss function $(x, y, t) \mapsto L_{\text{class}}(y, t)$, and hence we usually do not distinguish between L_{class} and its associated loss function. Since this kind of identification also occurs in the following examples, we will later formalize it in Definition 2.7.

Since this example is rather typical for classification problems in which the goal is to *detect* certain objects or events, we now present a weighted version of the classification scenario above.

Example 2.5 (Weighted binary classification). Let $Y := \{-1, 1\}$ and $\alpha \in (0, 1)$. Then the **α-weighted classification loss** $L_{\alpha\text{-class}} : Y \times \mathbb{R} \to [0, \infty)$ is defined by

$$L_{\alpha\text{-class}}(y, t) := \begin{cases} 1 - \alpha & \text{if } y = 1 \text{ and } t < 0 \\ \alpha & \text{if } y = -1 \text{ and } t \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

for all $y \in Y$, $t \in \mathbb{R}$. Obviously we have $2L_{1/2\text{-class}} = L_{\text{class}}$, i.e., the standard binary classification scenario is a special case of the general weighted classification scenario. Now, given a probability measure P on $X \times Y$ and a measurable $f : X \to \mathbb{R}$, the $L_{\alpha\text{-class}}$-risk can be computed by

$$\mathcal{R}_{L_{\alpha\text{-class}},\mathrm{P}}(f) = (1 - \alpha) \int_{f<0} \eta \, d\mathrm{P}_X + \alpha \int_{f \geq 0} (1 - \eta) \, d\mathrm{P}_X \,,$$

where again $\eta(x) := \mathrm{P}(y = 1|x)$, $x \in X$. From this we easily conclude that f is a Bayes decision function if and only if $(\eta(x) - \alpha) \operatorname{sign} f(x) \geq 0$ for P_X-almost all $x \in X$. Finally, the Bayes $L_{\alpha\text{-class}}$-risk is

$$\mathcal{R}^*_{L_{\alpha\text{-class}},\mathrm{P}} = \int_X \min\bigl\{(1 - \alpha)\eta, \alpha(1 - \eta)\bigr\} \, d\mathrm{P}_X \,. \qquad \triangleleft$$

In the two examples above the goal was to predict labels y from the set $\{-1, 1\}$. In the next example, we wish to predict general real-valued labels.

Example 2.6 (Least squares regression). The informal goal in regression is to predict the label $y \in Y := \mathbb{R}$ of a pair (x, y) drawn from an unknown probability measure P on $X \times Y$ if only x is observed. The most common way to formalize this goal is based on the **least squares loss** $L_{\mathrm{LS}} : Y \times \mathbb{R} \to [0, \infty)$ defined by

$$L_{\mathrm{LS}}(y, t) := (y - t)^2 \,, \qquad y \in Y, \, t \in \mathbb{R} \,. \tag{2.4}$$

In other words, the least squares loss penalizes the discrepancy between y and t *quadratically*. Obviously, for a measurable function $f : X \to \mathbb{R}$, the L_{LS}-risk is

$$\mathcal{R}_{L_{\mathrm{LS}},\mathrm{P}}(f) = \int_X \int_Y \bigl(y - f(x)\bigr)^2 \, d\mathrm{P}(y|x) \, d\mathrm{P}_X(x) \,.$$

By minimizing the inner integral with respect to $f(x)$, we then see that f is a Bayes decision function if and only if $f(x)$ almost surely equals the expected Y-value in x, i.e., if and only if

$$f(x) = \mathbb{E}_{\mathrm{P}}(Y|x) := \int_Y y \, d\mathrm{P}(y|x) \tag{2.5}$$

for P_X-almost all $x \in X$. Moreover, plugging $x \mapsto \mathbb{E}_{\mathrm{P}}(Y|x)$ into $\mathcal{R}_{L_{\mathrm{LS}},\mathrm{P}}(\cdot)$ shows that the Bayes L_{LS}-risk is the average conditional Y-variance, i.e.,

$$\mathcal{R}^*_{L_{\mathrm{LS}},\mathrm{P}} = \int_X \mathbb{E}_{\mathrm{P}}(Y^2|x) - \big(\mathbb{E}_{\mathrm{P}}(Y|x)\big)^2 \, d\mathrm{P}_X(x) \,.$$

Finally, an elementary calculation shows that the *excess* L_{LS}-risk of $f : X \to \mathbb{R}$ is

$$\mathcal{R}_{L_{\mathrm{LS}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{LS}},\mathrm{P}} = \int_X \big(\mathbb{E}_{\mathrm{P}}(Y|x) - f(x)\big)^2 d\mathrm{P}_X(x) \,,$$

i.e., if $\mathcal{R}_{L_{\mathrm{LS}},\mathrm{P}}(f)$ is close to $\mathcal{R}^*_{L_{\mathrm{LS}},\mathrm{P}}$, then f is close to the Bayes decision function in the sense of the $\|\cdot\|_{L_2(\mathrm{P}_X)}$. ◁

Using the least squares loss to make the informal regression goal precise seems to be rather arbitrary since, for example, for $p > 0$, the loss function

$$(y,t) \mapsto |y-t|^p \,, \qquad\qquad y \in \mathbb{R},\, t \in \mathbb{R} \,,$$

reflects the informal regression goal just as well. Nevertheless, the least squares loss is often chosen since it *"simplifies the mathematical treatment (and)... leads naturally to estimates which can be computed rapidly"*, as Györfi *et al.* (2002) write on p. 2. For SVMs, however, we will see later that none of these properties is exclusive for the least squares loss, and therefore we do not have to stick to the least squares loss for, e.g., computational reasons. On the other hand, the least squares loss is (essentially) the only loss whose Bayes decision functions have the form (2.5) for *all* distributions P with finite Bayes risk (see Proposition 3.44 for details), and hence the least squares loss is often the first choice when we wish to estimate the conditional expectations $\mathbb{E}_{\mathrm{P}}(Y|x)$, $x \in X$. Unfortunately, however, we will see in Chapter 10 that SVMs based on the least squares loss are rather sensitive to large deviations in y, and hence other losses may be preferred in some situations. We will discuss these questions in more detail in Sections 3.7 and 3.9 and Chapter 9.

A common feature of the loss functions above is that they are all *independent* of the input value x. Since this will also be the case for many other loss functions considered later, we introduce the following notion.

Definition 2.7. *A function $L : Y \times \mathbb{R} \to [0,\infty)$ is called a* ***supervised loss function****, or simply a* ***supervised loss****, if it is measurable.*

Note that a supervised loss L can be canonically identified with the loss function $\bar{L} : (x,y,t) \mapsto L(y,t)$. As in the examples above, we will thus write $\mathcal{R}_{L,\mathrm{P}}(f) := \mathcal{R}_{\bar{L},\mathrm{P}}(f)$ and $\mathcal{R}^*_{L,\mathrm{P}} := \mathcal{R}^*_{\bar{L},\mathrm{P}}$ in order to avoid notational overload.

Formally, we can also consider losses that are independent of y, i.e., we can introduce the following notion.

Definition 2.8. *A function $L : X \times \mathbb{R} \to [0,\infty)$ is called an* ***unsupervised loss function****, or simply an* ***unsupervised loss****, if it is measurable.*

Obviously, an unsupervised loss L can be canonically identified with the loss function $\bar{L} : (x, y, t) \mapsto L(x, t)$. As for supervised losses, we thus write

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

and $\mathcal{R}^*_{L,\mathrm{P}} := \mathcal{R}^*_{\bar{L},\mathrm{P}}$. Note that, in contrast to the risks for supervised losses, the risks for unsupervised losses are *independent of the supervisor* $\mathrm{P}(\,\cdot\,|x)$ that generates the labels. This explains the term "unsupervised loss". Since unsupervised losses do not depend on labeling information, these loss functions often occur in learning scenarios that lack labels in the available sample data. The two most important scenarios of this type are introduced in the following examples.

Example 2.9 (Density level detection). Let us suppose that we have some samples $D := (x_1, \ldots, x_n) \in X^n$ drawn in an i.i.d. fashion from an *unknown* distribution Q on X. Moreover, assume that our informal learning goal is to find the region where Q has relatively high concentration.

One way to formalize this learning goal is to assume that Q is absolutely continuous with respect to some known *reference measure* μ. Let $g : X \to [0, \infty)$ be the corresponding *unknown* density with respect to μ, i.e., $\mathrm{Q} = g\mu$. Then Q is highly concentrated in exactly the region where g is "large", i.e., our informal learning goal is to find the **density level sets** $\{g > \rho\}$ or $\{g \geq \rho\}$ for some fixed threshold $\rho > 0$. In order to find a formal specification of this learning goal, let us consider the unsupervised **density level detection (DLD) loss** $L_{\mathrm{DLD}} : X \times \mathbb{R} \to [0, \infty)$, which is defined by

$$L_{\mathrm{DLD}}(x, t) := \mathbf{1}_{(-\infty,0)}\big((g(x) - \rho)\operatorname{sign} t\big)\,, \qquad x \in X,\, t \in \mathbb{R}\,. \tag{2.6}$$

Note that for $f : X \to \mathbb{R}$ the loss $L_{\mathrm{DLD}}(x, f(x))$ penalizes the prediction $f(x)$ at x if either $f(x) \geq 0$ and $g(x) < \rho$, or $f(x) < 0$ and $g(x) > \rho$, whereas it ignores $f(x)$ if $g(x) = \rho$. In this sense, $\{f \geq 0\}$ is the prediction of f for our desired level set. In order to further formalize our informal learning goal, recall that the risks of unsupervised losses only depend on the marginal distributions P_X. In the density level detection scenario, we are mainly interested in the case $\mathrm{P}_X = \mu$, and thus we usually use the notation

$$\mathcal{R}_{L_{\mathrm{DLD}},\mu}(f) := \mathcal{R}_{L_{\mathrm{DLD}},\mathrm{P}}(f) = \int_X L_{\mathrm{DLD}}\big(x, f(x)\big)\, d\mu(x)\,.$$

From (2.6) it is then easy to conclude that a measurable $f : X \to \mathbb{R}$ is a Bayes decision function with respect to μ if and only if $\{g > \rho\} \subset \{f \geq 0\} \subset \{g \geq \rho\}$ holds true up to μ-zero sets. Consequently, we always have $\mathcal{R}^*_{L_{\mathrm{DLD}},\mu} = 0$ and, in addition, if $\mu(\{g = \rho\}) = 0$, we find

$$\mathcal{R}_{L_{\mathrm{DLD}},\mu}(f) = \mu\big(\{g \geq \rho\} \triangle \{f \geq 0\}\big)$$

for all measurable $f : X \to \mathbb{R}$, where $\triangle$ denotes the **symmetric difference** $A \triangle B := A \setminus B \cup B \setminus A$. In this sense, $\mathcal{R}_{L_{\mathrm{DLD}},\mu}(f)$ measures how well $\{f \geq 0\}$ detects the level set $\{g \geq \rho\}$.

Finally, observe that, unlike for the supervised loss functions of the previous examples, we *cannot compute* $L_{\mathrm{DLD}}(x,t)$ since g is unknown to us. Consequently, we cannot use an ERM scheme based on L_{DLD} simply because we cannot compute the empirical risk $\mathcal{R}_{L_{\mathrm{DLD}},\mathrm{D}}(f)$ for any $f : X \to \mathbb{R}$. Moreover, note that for the same reason we cannot estimate the quality of a found approximation $\{f \geq 0\}$ by $\mathcal{R}_{L_{\mathrm{DLD}},\mathrm{D}}(f)$ either. Because of these disadvantages of L_{DLD}, we will investigate more accessible supervised *surrogate* losses for L_{DLD} in Section 3.8. ◁

The density level detection scenario is often used if one wants to identify *anomalous future samples* $x \in X$ on the basis of unlabeled training data $D := (x_1, \ldots, x_n) \in X^n$. To this end, it is assumed that anomalous samples are somewhat atypical in the sense that they are not clustered. In other words, they occur in regions with low concentration, and consequently they are described by a level set $\{g \geq \rho\}$ for some suitably specified ρ.

In some sense, the density level detection scenario is an unsupervised counterpart of binary classification, and in fact we will establish a precise connection between these two in Section 3.8. The following, last example describes in a similar way an unsupervised counterpart of the regression scenario.

Example 2.10 (Density estimation). Let μ be a known probability measure on X and $g : X \to [0,\infty)$ be an *unknown* density with respect to μ. Let us further assume that our goal is to estimate the density g. Then one possible way to specify this goal is based on the unsupervised loss $L_q : X \times \mathbb{R} \to [0,\infty)$, $q > 0$, defined by

$$L_q(x,t) := \big|g(x) - t\big|^q, \qquad x \in X,\, t \in \mathbb{R}. \tag{2.7}$$

As for the DLD problem, we are usually interested in distributions P with $\mathrm{P}_X = \mu$, and for such we have

$$\mathcal{R}_{L_q,\mathrm{P}}(f) = \int_X \big|g(x) - f(x)\big|^q \, d\mu(x)$$

for all measurable $f : X \to \mathbb{R}$. From this we find $\mathcal{R}^*_{L_q,\mathrm{P}} = 0$ and, in addition, it is not hard to see that every Bayes decision function equals g modulo some μ-zero set. ◁

The presented examples of unsupervised learning scenarios suggest that the absence of labels is characteristic for situations where unsupervised losses occur. However, we will see in Chapter 3 that unsupervised losses are also a very powerful tool for investigating certain questions related to supervised learning scenarios.

2.2 Basic Properties of Loss Functions and Their Risks

In this section, we introduce some additional features of loss functions such as convexity, continuity, and differentiability and relate these features to analogous features of the associated risks. Since the results of this section will be used throughout this book, we recommend that even the experienced reader becomes familiar with the material of this section.

Our first lemma shows that under some circumstances risk functionals are measurable.

Lemma 2.11 (Measurability of risks). *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss and $\mathcal{F} \subset \mathcal{L}_0(X)$ be a subset that is equipped with a complete and separable metric d and its corresponding Borel σ-algebra. Assume that the metric d* ***dominates the pointwise convergence****, i.e.,*

$$\lim_{n\to\infty} d(f_n, f) = 0 \qquad \Longrightarrow \qquad \lim_{n\to\infty} f_n(x) = f(x)\,, \qquad x \in X, \tag{2.8}$$

for all $f, f_n \in \mathcal{F}$. Then the evaluation map

$$\begin{aligned} \mathcal{F} \times X &\to \mathbb{R} \\ (f, x) &\mapsto f(x) \end{aligned}$$

is measurable, and consequently the map $(x, y, f) \mapsto L(x, y, f(x))$ defined on $X \times Y \times \mathcal{F}$ is also measurable. Finally, given a distribution P on $X \times Y$, the risk functional $\mathcal{R}_{L,\mathrm{P}} : \mathcal{F} \to [0, \infty]$ is measurable.

Proof. Since d dominates the pointwise convergence, we see that, for fixed $x \in X$, the $\mathbb{R}$-valued map $f \mapsto f(x)$ defined on $\mathcal{F}$ is continuous with respect to d. Furthermore, $\mathcal{F} \subset \mathcal{L}_0(X)$ implies that, for fixed $f \in \mathcal{F}$, the $\mathbb{R}$-valued map $x \mapsto f(x)$ defined on X is measurable. By Lemma A.3.17, we then obtain the first assertion. Since this implies that the map $(x, y, f) \mapsto (x, y, f(x))$ is measurable, we obtain the second assertion. The third assertion now follows from the measurability statement in Tonelli's Theorem A.3.10. □

Obviously, the metric defined by the supremum norm $\|\cdot\|_\infty$ dominates the pointwise convergence for every $\mathcal{F} \subset \mathcal{L}_\infty(X)$. Moreover, we will see in Section 4.2 that the metric of reproducing kernel Hilbert spaces also dominates the pointwise convergence.

Let us now consider some additional properties of loss functions and their risks. We begin with convexity.

Definition 2.12. *A loss $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is called* ***(strictly) convex*** *if $L(x, y, \cdot) : \mathbb{R} \to [0, \infty)$ is (strictly) convex for all $x \in X$ and $y \in Y$.*

If L is a supervised or unsupervised loss function, then we call L (strictly) convex if its canonically associated loss function $\bar{L}$ is (strictly) convex. In the

following, we will analogously assign other properties to L via its identification with $\bar{L}$.

The next simple lemma, whose proof is left as an exercise, shows that convexity of the loss implies convexity of its risks.

Lemma 2.13 (Convexity of risks). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a (strictly) convex loss and P be a distribution on $X \times Y$. Then $\mathcal{R}_{L,\mathrm{P}} : \mathcal{L}_0(X) \to [0,\infty]$ is (strictly) convex.*

Besides convexity we also need some notions of continuity for loss functions. We begin with a qualitative definition.

Definition 2.14. *A loss $L : X \times Y \times \mathbb{R} \to [0,\infty)$ is called **continuous** if $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ is continuous for all $x \in X$, $y \in Y$.*

If we have a continuous loss function $L : X \times Y \times \mathbb{R} \to [0,\infty)$ and a sequence (f_n) of measurable functions $f_n : X \to \mathbb{R}$ that converges pointwise to a function $f : X \to \mathbb{R}$, then we obviously have $L(x,y,f_n(x)) \to L(x,y,f(x))$ for all $(x,y) \in X \times Y$. However, it is well-known from integration theory that such a convergence does *not* imply a convergence of the corresponding integrals, i.e., in general we cannot conclude $\mathcal{R}_{L,\mathrm{P}}(f_n) \to \mathcal{R}_{L,\mathrm{P}}(f)$. However, the following, weaker result always holds.

Lemma 2.15 (Lower semi-continuity of risks). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a continuous loss, P be a distribution on $X \times Y$, and $(f_n) \subset \mathcal{L}_0(\mathrm{P}_X)$ be a sequence that converges to an $f \in \mathcal{L}_0(\mathrm{P}_X)$ in probability with respect to the marginal distribution P_X. Then we have*

$$\mathcal{R}_{L,\mathrm{P}}(f) \leq \liminf_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n)\,.$$

Proof. Since (f_n) converges in probability P_X, there exists a subsequence (f_{n_k}) of (f_n) with

$$\lim_{k\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_{n_k}) = \liminf_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n)$$

and $f_{n_k}(x) \to f(x)$ for P_X-almost all $x \in X$. By the continuity of L, we then have $L(x,y,f_{n_k}(x)) \to L(x,y,f(x))$ for P-almost all $(x,y) \in X \times Y$, and hence Fatou's lemma (see Theorem A.3.4) gives

$$\begin{aligned}
\mathcal{R}_{L,\mathrm{P}}(f) &= \int_{X\times Y} \lim_{k\to\infty} L\bigl(x,y,f_{n_k}(x)\bigr)\, d\mathrm{P}(x,y) \\
&\leq \liminf_{k\to\infty} \int_{X\times Y} L\bigl(x,y,f_{n_k}(x)\bigr)\, d\mathrm{P}(x,y) \\
&= \liminf_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n)\,.
\end{aligned}$$

□

If we have an integrable majorant of the sequence $L(\cdot,\cdot,f_n(\cdot))$ in the proof above, Lebesgue's Theorem A.3.6 obviously gives $\mathcal{R}_{L,\mathrm{P}}(f_n) \to \mathcal{R}_{L,\mathrm{P}}(f)$. The following definition describes losses for which we have such a majorant.

Definition 2.16. *We call a loss $L : X \times Y \times \mathbb{R} \to [0,\infty)$ a* ***Nemitski loss*** *if there exist a measurable function $b : X \times Y \to [0,\infty)$ and an increasing function $h : [0,\infty) \to [0,\infty)$ such that*

$$L(x,y,t) \le b(x,y) + h(|t|)\,, \qquad (x,y,t) \in X \times Y \times \mathbb{R}. \tag{2.9}$$

Furthermore, we say that L is a ***Nemitski loss of order*** *$p \in (0,\infty)$ if there exists a constant $c > 0$ such that*

$$L(x,y,t) \le b(x,y) + c\,|t|^p\,, \qquad (x,y,t) \in X \times Y \times \mathbb{R}.$$

Finally, if P is a distribution on $X \times Y$ with $b \in \mathcal{L}_1(\mathrm{P})$, we say that L is a ***P-integrable Nemitski loss.***

Note that P-integrable Nemitski losses L satisfy $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$ for all $f \in L_\infty(\mathrm{P}_X)$, and consequently we also have $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$ and $\mathcal{R}^*_{L,\mathrm{P}} < \infty$. In addition, we should keep in mind that the notion of Nemitski losses will become of particular interest when dealing with *unbounded* Y, which is typical for the regression problems treated in Chapters 9 and 10.

Let us now investigate the continuity of risks based on Nemitski losses.

Lemma 2.17 (Continuity of risks). *Let P be a distribution on $X \times Y$ and $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a continuous, P-integrable Nemitski loss. Then the following statements hold:*

i) Let $f_n : X \to \mathbb{R}$, $n \ge 1$, be bounded, measurable functions for which there exists a constant $B > 0$ with $\|f_n\|_\infty \le B$ for all $n \ge 1$. If the sequence (f_n) converges P_X-almost surely to a function $f : X \to \mathbb{R}$, then we have

$$\lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n) = \mathcal{R}_{L,\mathrm{P}}(f)\,.$$

ii) The map $\mathcal{R}_{L,\mathrm{P}} : L_\infty(\mathrm{P}_X) \to [0,\infty)$ is well-defined and continuous.

iii) If L is of order $p \in [1,\infty)$, then $\mathcal{R}_{L,\mathrm{P}} : L_p(\mathrm{P}_X) \to [0,\infty)$ is well-defined and continuous.

Proof. i). Obviously, f is a bounded and measurable function with $\|f\|_\infty \le B$. Furthermore, the continuity of L shows

$$\lim_{n\to\infty} \bigl|L(x,y,f_n(x)) - L(x,y,f(x))\bigr| = 0$$

for P-almost all $(x,y) \in X \times Y$. In addition, we have

$$\begin{aligned}\bigl|L(x,y,f_n(x)) - L(x,y,f(x))\bigr| &\le 2b(x,y) + h(|f_n(x)|) + h(|f(x)|)\\ &\le 2b(x,y) + 2h(B)\end{aligned}$$

for all $(x,y) \in X \times Y$ and all $n \ge 1$. Since the function $2b(\,\cdot\,,\,\cdot\,) + 2h(B)$ is P-integrable, Lebesgue's theorem together with

$$\bigl|\mathcal{R}_{L,\mathrm{P}}(f_n) - \mathcal{R}_{L,\mathrm{P}}(f)\bigr| \leq \int_{X\times Y} \bigl|L(x,y,f_n(x)) - L(x,y,f(x))\bigr|\, d\mathrm{P}(x,y)$$

gives the assertion.

ii). Condition (2.9) together with $b \in \mathcal{L}_1(\mathrm{P})$ obviously ensures $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$ for all $f \in L_\infty(\mathrm{P}_X)$, i.e., $\mathcal{R}_{L,\mathrm{P}}$ actually maps $L_\infty(\mathrm{P}_X)$ into $[0,\infty)$. Moreover, the continuity is a direct consequence of *i).*

iii). Since L is a P-integrable Nemitski loss of order p, we obviously have $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$ for all $f \in L_p(\mathrm{P}_X)$. Now let $(f_n) \subset L_p(\mathrm{P}_X)$ be a convergent sequence with limit $f \in L_p(\mathrm{P}_X)$. Since convergence in $L_p(\mathrm{P}_X)$ implies convergence in probability P_X, Lemma 2.15 then yields

$$\mathcal{R}_{L,\mathrm{P}}(f) \leq \liminf_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n)\,.$$

Moreover, $\tilde{L}(x,y,t) := b(x,y) + c\,|t|^p - L(x,y,t)$ defines a continuous loss function, and hence Lemma 2.15 also gives

$$\begin{aligned}
\|b\|_{L_1(\mathrm{P})} + c\|f\|_p^p - \mathcal{R}_{L,\mathrm{P}}(f) &= \mathcal{R}_{\tilde{L},\mathrm{P}}(f) \\
&\leq \liminf_{n\to\infty} \mathcal{R}_{\tilde{L},\mathrm{P}}(f_n) \\
&= \liminf_{n\to\infty} \bigl(\|b\|_{L_1(\mathrm{P})} + c\|f_n\|_p^p - \mathcal{R}_{L,\mathrm{P}}(f_n)\bigr)\,.
\end{aligned}$$

Using that $\|\cdot\|_p^p$ is continuous on $L_p(\mathrm{P}_X)$, we thus obtain

$$\limsup_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n) \leq \mathcal{R}_{L,\mathrm{P}}(f)\,. \qquad \square$$

Let us now turn to a quantitative notion of continuity for loss functions.

Definition 2.18. *A loss $L : X \times Y \times \mathbb{R} \to [0,\infty)$ is called **locally Lipschitz continuous** if for all $a \geq 0$ there exists a constant $c_a \geq 0$ such that*

$$\sup_{\substack{x\in X\\ y\in Y}} \bigl|L(x,y,t) - L(x,y,t')\bigr| \;\leq\; c_a\,|t-t'|\,, \qquad t,t' \in [-a,a]\,. \tag{2.10}$$

*Moreover, for $a \geq 0$, the smallest such constant c_a is denoted by $|L|_{a,1}$. Finally, if we have $|L|_1 := \sup_{a\geq 0} |L|_{a,1} < \infty$, we call L **Lipschitz continuous**.*

Note that if Y is *finite* and $L : Y \times \mathbb{R} \to [0,\infty)$ is a *supervised* convex loss, then L is locally Lipschitz continuous since every convex function is locally Lipschitz continuous by Lemma A.6.5. Furthermore, a locally Lipschitz continuous loss L is a Nemitski loss since the definition of $|L|_{|t|,1}$ yields

$$L(x,y,t) \leq L(x,y,0) + |L|_{|t|,1}|t|\,, \qquad (x,y) \in X \times Y,\, t \in \mathbb{R}. \tag{2.11}$$

In particular, a locally Lipschitz continuous loss L is a P-integrable Nemitski loss if and only if $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$. Finally, if L is Lipschitz continuous, then L is a Nemitski loss of order $p = 1$.

The following lemma, whose proof is left as an exercise, relates the (local) Lipschitz continuity of L to the (local) Lipschitz continuity of its risk.

Lemma 2.19 (Lipschitz continuity of risks). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a locally Lipschitz continuous loss and P be a distribution on $X \times Y$. Then for all $B \geq 0$ and all $f, g \in L_\infty(\mathrm{P}_X)$ with $\|f\|_\infty \leq B$ and $\|g\|_\infty \leq B$, we have*

$$\bigl|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(g)\bigr| \;\leq\; |L|_{B,1} \cdot \|f - g\|_{L_1(\mathrm{P}_X)} .$$

Our next goal is to consider the differentiability of risks. To this end, we first have to introduce differentiable loss functions in the following definition.

Definition 2.20. *A loss $L : X \times Y \times \mathbb{R} \to [0,\infty)$ is called **differentiable** if $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ is differentiable for all $x \in X$, $y \in Y$. In this case, $L'(x,y,t)$ denotes the derivative of $L(x,y,\cdot)$ at $t \in \mathbb{R}$.*

In general, we cannot expect that the risk of a differentiable loss function is differentiable. However, for certain integrable Nemitski losses, we can actually establish the differentiability of the associated risk.

Lemma 2.21 (Differentiability of risks). *Let P be a distribution on $X \times Y$ and $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a differentiable loss such that both L and $|L'| : X \times Y \times \mathbb{R} \to [0,\infty)$ are P-integrable Nemitski losses. Then the risk functional $\mathcal{R}_{L,\mathrm{P}} : L_\infty(\mathrm{P}_X) \to [0,\infty)$ is Fréchet differentiable and its derivative at $f \in L_\infty(\mathrm{P}_X)$ is the bounded linear operator $\mathcal{R}'_{L,\mathrm{P}}(f) : L_\infty(\mathrm{P}_X) \to \mathbb{R}$ given by*

$$\mathcal{R}'_{L,\mathrm{P}}(f)g = \int_{X\times Y} g(x) L'\bigl(x,y,f(x)\bigr)\, d\mathrm{P}(x,y)\,, \qquad g \in L_\infty(\mathrm{P}_X).$$

Proof. We first observe that the derivative $L' : X \times Y \times \mathbb{R} \to \mathbb{R}$ is measurable since

$$L'(x,y,t) = \lim_{n\to\infty} \frac{L(x,y,t+1/n) - L(x,y,t)}{1/n}\,, \qquad (x,y,t) \in X \times Y \times \mathbb{R}.$$

Now let $f \in L_\infty(\mathrm{P}_X)$ and $(f_n) \subset L_\infty(\mathrm{P}_X)$ be a sequence with $f_n \neq 0$, $n \geq 1$, and $\lim_{n\to\infty} \|f_n\|_\infty = 0$. Without loss of generality, we additionally assume for later use that $\|f_n\|_\infty \leq 1$ for all $n \geq 1$. For $(x,y) \in X \times Y$ and $n \geq 1$, we now define

$$G_n(x,y) := \left| \frac{L\bigl(x,y,f(x)+f_n(x)\bigr) - L\bigl(x,y,f(x)\bigr)}{f_n(x)} - L'\bigl(x,y,f(x)\bigr) \right|$$

if $f_n(x) \neq 0$, and $G_n(x,y) := 0$ otherwise. Now an easy estimation gives

$$\begin{aligned}
&\left| \frac{\mathcal{R}_{L,\mathrm{P}}(f+f_n) - \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}'_{L,\mathrm{P}}(f) f_n}{\|f_n\|_\infty} \right| \\
&\leq \int\limits_{X\times Y} \left| \frac{L\bigl(x,y,f(x)+f_n(x)\bigr) - L\bigl(x,y,f(x)\bigr) - f_n(x) L'\bigl(x,y,f(x)\bigr)}{\|f_n\|_\infty} \right| d\mathrm{P}(x,y) \\
&\leq \int\limits_{X\times Y} G_n(x,y)\, d\mathrm{P}(x,y) \qquad (2.12)
\end{aligned}$$

for all $n \geq 1$. Furthermore, for $(x, y) \in X \times Y$, the definitions of G_n and $L'(x, y, \cdot)$ obviously yield

$$\lim_{n\to\infty} G_n(x, y) = 0. \tag{2.13}$$

Moreover, for $(x, y) \in X \times Y$ and $n \geq 1$ with $f_n(x) \neq 0$, the mean value theorem shows that there exists a $g_n(x, y)$ with $|g_n(x, y)| \in [0, |f_n(x)|]$ and

$$\frac{L\big(x, y, f(x) + f_n(x)\big) - L\big(x, y, f(x)\big)}{f_n(x)} = L'\big(x, y, f(x) + g_n(x, y)\big)\,.$$

Since $|L'|$ is a P-integrable Nemitski loss, there also exist a $b : X \times Y \to [0, \infty)$ with $b \in L_1(\mathrm{P})$ and an increasing function $h : [0, \infty) \to [0, \infty)$ with

$$|L'(x, y, t)| \leq b(x, y) + h(t)\,, \qquad (x, y, t) \in X \times Y \times \mathbb{R}.$$

This together with $\|f_n\|_\infty \leq 1$, $n \geq 1$, implies

$$\begin{aligned}\left|\frac{L\big(x, y, f(x) + f_n(x)\big) - L\big(x, y, f(x)\big)}{f_n(x)}\right| &\leq b(x, y) + h\big(|f(x) + g_n(x, y)|\big)\\ &\leq b(x, y) + h\big(\|f\|_\infty + 1\big)\end{aligned}$$

for all $(x, y) \in X \times Y$ and $n \geq 1$ with $f_n(x) \neq 0$. Since these estimates show that

$$G_n(x, y) \leq 2b(x, y) + 2h\big(\|f\|_\infty + 1\big)$$

for all $(x, y) \in X \times Y$ and $n \geq 1$, we then obtain the assertion by (2.12), (2.13), and Lebesgue's Theorem A.3.6. □

Our last goal in this section is to investigate loss functions that in some sense can be restricted to domains of the form $X \times Y \times [-M, M]$.

Definition 2.22. *We say that a loss $L : X \times Y \times \mathbb{R} \to [0, \infty)$ can be **clipped** at $M > 0$ if, for all $(x, y, t) \in X \times Y \times \mathbb{R}$, we have*

$$L(x, y, \widehat{t}\,) \;\leq\; L(x, y, t)\,,$$

*where $\widehat{t}$ denotes the **clipped** value of t at $\pm M$, that is*

$$\widehat{t} := \begin{cases} -M & \text{if } t < -M \\ t & \text{if } t \in [-M, M] \\ M & \text{if } t > M\,. \end{cases} \tag{2.14}$$

Moreover, we say that L can be clipped if it can be clipped at some $M > 0$.

For most losses, it is elementary to check whether they can be clipped, but for convex losses this work can be further simplified by the following elementary criterion.

Lemma 2.23 (Clipped convex losses). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss and $M > 0$. Then the following statements are equivalent:*

i) L can be clipped at M.
ii) For all $(x,y) \in X \times Y$, the function $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ has at least one global minimizer in $[-M,M]$.

Proof. For $(x,y) \in X \times Y$, we denote the set of minimizers of $L(x,y,\cdot)$ by $\mathcal{M}_{x,y} := \{t^* \in \mathbb{R} : L(x,y,t^*) = \inf_{t\in\mathbb{R}} L(x,y,t)\}$. For later use, note that the convexity of L implies that $\mathcal{M}_{x,y}$ is a closed interval by Lemma A.6.2.

i) $\Rightarrow$ *ii)*. Assume that there exists a pair $(x,y) \in X \times Y$ such that $\mathcal{M}_{x,y} \cap [-M,M] = \emptyset$. In the case $\mathcal{M}_{x,y} = \emptyset$, the convexity of L shows that $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ is strictly monotone and hence L cannot be clipped at any real number. Therefore we may assume without loss of generality that $t := \inf \mathcal{M}_{x,y}$ satisfies $M < t < \infty$. However, in this case we have

$$L(x,y,\widehat{t}) = L(x,y,M) > L(x,y,t)\,,$$

i.e., L cannot be clipped at M.

ii) $\Rightarrow$ *i)*. Our assumption *ii)* guarantees $\mathcal{M}_{x,y} \cap [-M,M] \neq \emptyset$, and hence we have $\inf \mathcal{M}_{x,y} \leq M$ and $\sup \mathcal{M}_{x,y} \geq -M$. Moreover, the convexity of L shows that $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ is increasing on $[\sup \mathcal{M}_{x,y},\infty)$ and decreasing on $(-\infty, \inf \mathcal{M}_{x,y}]$, and hence L can be clipped at M. □

The criterion above will be of particular interest in Section 7.4, where we investigate the statistical properties of SVMs that use clippable losses. Therefore, it will be important to remember that, for the loss functions introduced in the following sections, condition *ii)* is usually elementary to check.

2.3 Margin-Based Losses for Classification Problems

In Examples 2.4 and 2.5, we considered the (weighted) binary classification scenario, which is described by the supervised loss functions L_{class} and $L_{\alpha\text{-class}}$, respectively. Now observe that both loss functions are not *convex*, which may lead to computational problems if, for example, one tries to minimize an empirical risk $\mathcal{R}_{L_{\text{class}},D}(\cdot)$ over some set $\mathcal{F}$ of functions $f : X \to \mathbb{R}$. This is the reason why many machine learning algorithms consider the empirical risk $\mathcal{R}_{L,D}(\cdot)$ of a *surrogate* loss function $L : Y \times \mathbb{R} \to [0,\infty)$ instead. In this section, we will introduce some commonly used surrogate losses and establish a few basic properties of these losses. Finally, we show why the hinge loss used in SVMs for classification is a good surrogate.

Throughout this section, we assume $Y := \{-1,1\}$. Let us begin with the following basic definition, which introduces a type of loss function often used in classification algorithms.

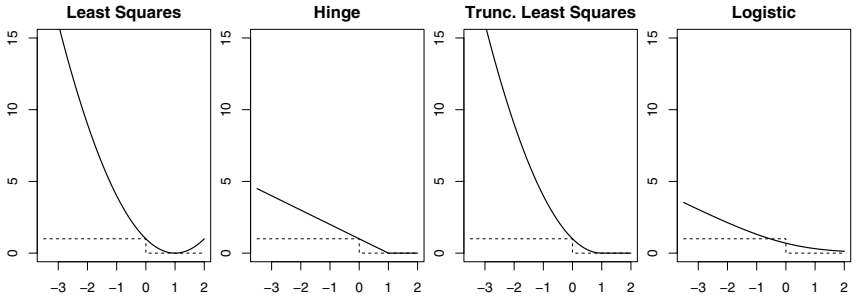

Fig. 2.1. The shape of the representing function φ for some margin-based loss functions considered in the text.

Definition 2.24. *A supervised loss $L : Y \times \mathbb{R} \to [0, \infty)$ is called* ***margin-based*** *if there exists a* ***representing function*** *$\varphi : \mathbb{R} \to [0, \infty)$ such that*

$$L(y, t) = \varphi(yt)\,, \qquad y \in Y,\, t \in \mathbb{R}.$$

The following lemma relates some simple properties of margin-based losses to analogous properties of their representing functions.

Lemma 2.25 (Properties of margin-based losses). *Let L be a margin-based loss represented by φ. Then the following statements are true:*

i) L is (strictly) convex if and only if φ is (strictly) convex.
ii) L is continuous if and only if φ is.
iii) L is (locally) Lipschitz continuous if and only if φ is.
iv) If L is convex, then it is locally Lipschitz continuous.
v) L is a P-integrable Nemitski loss for all measurable spaces X and all distributions P on $X \times Y$.

Proof. Recalling the definitions of Section 2.2, the first three assertions are trivial and *iv*) follows from Lemma A.6.5. Finally, *v)* follows from

$$L(y, t) \le \max\bigl\{\varphi(-t), \varphi(t)\bigr\}\,, \qquad y \in Y,\, t \in \mathbb{R}. \qquad \square$$

Note that the classification loss L_{class} is *not* margin-based, while many commonly used surrogates for L_{class} are margin-based. We are in particular interested in the following examples (see also Figure 2.1 for some illustrations).

Example 2.26. The **least squares loss** L_{LS} is margin-based since it satisfies

$$L_{\text{LS}}(y, t) = (y - t)^2 = (1 - yt)^2\,, \qquad y = \pm 1,\, t \in \mathbb{R}.$$

In addition, L_{LS} is obviously strictly convex, and for $a > 0$ its local Lipschitz constant is $|L_{\text{LS}}|_{a,1} = 2a + 2$ by Lemma A.6.8. Finally, L_{LS} can be clipped at $M = 1$. $\triangleleft$

Example 2.27. The **hinge loss** $L_{\text{hinge}} : Y \times \mathbb{R} \to [0, \infty)$ is defined by

$$L_{\text{hinge}}(y, t) := \max\{0, 1 - yt\}\,, \qquad y = \pm 1,\ t \in \mathbb{R}.$$

It is clearly a margin-based loss that linearly penalizes every prediction t with $yt \leq 1$. In addition, it is obviously convex and Lipschitz continuous with $|L_{\text{hinge}}|_1 = 1$. Finally, L_{hinge} can be clipped at $M = 1$. ◁

Example 2.28. The **truncated least squares loss** or **squared hinge loss** is defined by

$$L_{\text{trunc-ls}}(y, t) := \big(\max\{0, 1 - yt\}\big)^2\,, \qquad y = \pm 1,\ t \in \mathbb{R}.$$

It is obviously a margin-based loss that quadratically penalizes every prediction t with $yt \leq 1$. In addition, it is convex, and its local Lipschitz constants are $|L_{\text{trunc-ls}}|_{a,1} = 2a + 2$, $a > 0$. Finally, L_{LS} can be clipped at $M = 1$. ◁

Example 2.29. The **logistic loss for classification** $L_{\text{c-logist}}$ is defined by

$$L_{\text{c-logist}}(y, t) := \ln\big(1 + \exp(-yt)\big)\,, \qquad y = \pm 1,\ t \in \mathbb{R}.$$

It is obviously a margin-based loss function whose shape is close to that of the hinge loss. However, unlike the hinge loss, the logistic loss is infinitely many times differentiable. In addition, it is strictly convex and Lipschitz continuous with $|L_{\text{c-logist}}|_1 = 1$. Finally, $L_{\text{c-logist}}$ cannot be clipped. ◁

Let us finally investigate in which sense the hinge loss used in the soft margin SVM is a reasonable surrogate for the classification loss. To this end, we need the following elementary lemma.

Lemma 2.30. *For all $\eta \in [0, 1]$ and all $t \in [-1, 1]$, we have*

$$|2\eta - 1|\mathbf{1}_{(-\infty,0]}\big((2\eta - 1)\operatorname{sign} t\big) \;\leq\; |2\eta - 1| \cdot \big|t - \operatorname{sign}(2\eta - 1)\big|\,. \tag{2.15}$$

Proof. For $\eta = 1/2$, there is nothing to prove. In order to prove the other cases, let us first recall our convention $\operatorname{sign} 0 := 1$. For $\eta \in [0, 1/2)$ and $t \in [-1, 0)$, we now have $(2\eta - 1)\operatorname{sign} t > 0$, and hence the left-hand side of (2.15) equals zero. From this we immediately obtain the assertion. Moreover, for $t \in [0, 1]$, we have $(2\eta - 1)\operatorname{sign} t < 0$, which in turn yields

$$|2\eta - 1|\mathbf{1}_{(-\infty,0]}\big((2\eta - 1)\operatorname{sign} t\big) \leq |2\eta - 1| \cdot (t + 1) = |2\eta - 1| \cdot \big|t - \operatorname{sign}(2\eta - 1)\big|\,.$$

In other words, we have shown the assertion for $\eta < 1/2$. The case $\eta > 1/2$ can be shown completely analogously and is left as an additional exercise for the reader. □

With the help of the lemma above we can now investigate the relationship between the L_{hinge}-risk and the classification risk.

Theorem 2.31 (Zhang's inequality). *Given a distribution* P *on* $X \times Y$*, we write* $\eta(x) := \mathrm{P}(y = 1|x)$*,* $x \in X$*. Moreover, let* $f^*_{L_{\text{class}},\mathrm{P}}$ *be the Bayes classification function given by* $f^*_{L_{\text{class}},\mathrm{P}}(x) := \operatorname{sign}(2\eta(x) - 1)$*,* $x \in X$*. Then, for all measurable* $f : X \to [-1, 1]$*, we have*

$$\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} = \int_X |f(x) - f^*_{L_{\text{class}},\mathrm{P}}(x)| \cdot |2\eta(x) - 1| \, d\mathrm{P}_X(x) \,.$$

Moreover, for every measurable $f : X \to \mathbb{R}$*, we have*

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

Proof. For $f : X \to [-1, 1]$, the definition of the hinge loss yields

$$\begin{aligned}\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(f) &= \int_X (1 - f(x))\,\eta(x) + (1 + f(x))(1 - \eta(x))\, d\mathrm{P}_X(x) \\ &= \int_X 1 + f(x)(1 - 2\eta(x))\, d\mathrm{P}_X(x) \,,\end{aligned}$$

which in turn implies $\mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} = \mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(f^*_{L_{\text{class}},\mathrm{P}})$ since the hinge loss can be clipped at $M = 1$ by Lemma 2.23. From this we conclude that

$$\begin{aligned}\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} &= \int_X f(x)(1 - 2\eta(x)) + |2\eta(x) - 1| \, d\mathrm{P}_X(x) \\ &= \int_X |f(x) - f^*_{L_{\text{class}},\mathrm{P}}(x)| \cdot |2\eta(x) - 1| \, d\mathrm{P}_X(x) \,,\end{aligned}$$

i.e., we have shown the first assertion. To prove the second assertion, we first use that L_{hinge} can be clipped at $M = 1$ to obtain

$$\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(\widehat{f}\,) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} \;\le\; \mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}}$$

for the clipped version $\widehat{f}$ of a function $f : X \to \mathbb{R}$. Moreover, this clipped version also satisfies

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} = \mathcal{R}_{L_{\text{class}},\mathrm{P}}(\widehat{f}\,) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} \,,$$

and consequently it suffices to show the second assertion for $f : X \to [-1, 1]$. Now recall Example 2.4, where we saw $\mathcal{R}^*_{L_{\text{class}},\mathrm{P}} = \mathcal{R}_{L_{\text{class}},\mathrm{P}}(f^*_{L_{\text{class}},\mathrm{P}})$ and

$$\begin{aligned}&\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} \\ &= \int_X \eta\, \mathbf{1}_{(-\infty,0)}(f) + (1 - \eta)\, \mathbf{1}_{[0,\infty)}(f) - \min\{\eta, 1 - \eta\}\, d\mathrm{P}_X \\ &= \int_X |2\eta(x) - 1|\, \mathbf{1}_{(-\infty,0]}\big((2\eta(x) - 1) \operatorname{sign} f(x)\big)\, d\mathrm{P}_X(x) \,.\end{aligned}$$

Lemma 2.30 and the first assertion then yield the second assertion. □

Recall that the goal in binary classification was to find a function f whose excess classification risk $\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}}$ is small. By Theorem 2.31, we now see that we achieve this goal whenever $\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}}$ is small. In this sense, the hinge loss is a reasonable surrogate for the classification loss. Finally, note that we will show in Section 3.4 that the other margin-based losses introduced in this section are also reasonable surrogates.

Finally, observe that *all* calculations in the preceding proof are *solely* in terms of $\eta(x) = \mathrm{P}(y = 1|x)$ and $f(x)$. This observation will be the key trick for analyzing general surrogate losses in Chapter 3.

2.4 Distance-Based Losses for Regression Problems

In regression, the problem 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 we introduced in Example 2.6. However, we also mentioned there that this is by no means the only reasonable loss. In this section, we therefore introduce some other loss functions for the regression problem. In addition, we establish some basic properties of these losses and their associated risks.

Let us begin with the following basic definitions.

Definition 2.32. *We say that a supervised loss $L : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ is:*

i) ***distance-based*** *if there exists a* ***representing function*** *$\psi : \mathbb{R} \to [0, \infty)$ satisfying $\psi(0) = 0$ and*

$$L(y, t) = \psi(y - t)\,, \qquad y \in Y,\, t \in \mathbb{R};$$

ii) ***symmetric*** *if L is distance-based and its representing function ψ satisfies*

$$\psi(r) = \psi(-r)\,, \qquad r \in \mathbb{R}.$$

Obviously, the least squares loss as well as the family of losses mentioned after Example 2.6 are symmetric loss functions. Further examples of this type of loss will be presented later in this section. Let us first, however, establish some basic properties of distance-based losses and their associated risks. We begin with the following lemma, which relates properties of L with properties of ψ. Its proof is left as an exercise.

Lemma 2.33 (Properties of distance-based losses). *Let L be a distance-based loss with representing function $\psi : \mathbb{R} \to [0, \infty)$. Then we have:*

i) L is (strictly) convex if and only if ψ is (strictly) convex.
ii) L is continuous if and only if ψ is continuous.
iii) L is Lipschitz continuous if and only if ψ is Lipschitz continuous.

Note that the *local* Lipschitz continuity of ψ does *not* imply the local Lipschitz continuity of the corresponding distance-based loss function as, for example, the least squares loss shows.

Our next goal is to investigate under which conditions on the distribution P a distance-based loss function is a P-integrable Nemitski loss. This analysis will be conducted in two steps: *a)* the analysis of the integrals of the form

$$\mathcal{C}_{L,\mathrm{Q}}(t) := \int_{\mathbb{R}} L(y,t)\, d\mathrm{Q}(y)\,, \tag{2.16}$$

which occur for $\mathrm{Q} := \mathrm{P}(Y|x)$ as inner integrals in the definition of the L-risk, and *b)* a subsequent analysis of the averaging with respect to P_X. For the first step, we need the following definition, which will be used to describe the tail behavior of the conditional distributions $\mathrm{P}(Y|x)$.

Definition 2.34. *For a distribution* Q *on* $\mathbb{R}$*, the* p***-th moment****,* $p \in (0,\infty)$*, is defined by*

$$|\mathrm{Q}|_p := \left(\int_{\mathbb{R}} |y|^p \, d\mathrm{Q}(y) \right)^{1/p}.$$

Moreover, its ∞***-moment*** *is defined by* $|\mathrm{Q}|_\infty := \sup\bigl|\operatorname{supp} Q\bigr|$.

Note that in general the p-th moment of a distribution Q on $\mathbb{R}$ is not finite. In particular, we have $|\mathrm{Q}|_\infty < \infty$ if and only if Q has a bounded support. Moreover, for $0 < p \le q \le \infty$, we always have $|\mathrm{Q}|_p \le |\mathrm{Q}|_q$.

Besides controlling the tail behavior of the conditional distributions we also need to describe the growth behavior of the loss function considered. This is done in the following definition.

Definition 2.35. *Let* $p \in (0,\infty)$ *and* $L : \mathbb{R} \times \mathbb{R} \to [0,\infty)$ *be a distance-based loss with representing function* ψ*. We say that* L *is of:*

i) ***upper growth*** p *if there is a constant* $c > 0$ *such that*

$$\psi(r) \le c\left(|r|^p + 1\right), \qquad r \in \mathbb{R};$$

ii) ***lower growth*** p *if there is a constant* $c > 0$ *such that*

$$\psi(r) \ge c\left(|r|^p - 1\right), \qquad r \in \mathbb{R};$$

iii) ***growth type*** p *if* L *is of both upper and lower growth type* p*.*

Our next goal is to relate the tail behavior of the conditional distributions with the growth behavior of L and the integrals (2.16). To this end, recall that convex functions are locally Lipschitz continuous (see Lemma A.6.5), and hence, for *convex* distance-based loss functions L, the representing ψ is locally Lipschitz continuous on every interval $[-r,r]$. Consequently,

$$r \mapsto |\psi_{|[-r,r]}|_1\,, \qquad r \ge 0, \tag{2.17}$$

defines an increasing, non-negative function. The following lemma establishes some basic properties of this function and relates them to the growth type of distance-based loss functions.

Lemma 2.36 (Growth type and moments). *Let L be a distance-based loss with representing function ψ and Q be a distribution on $\mathbb{R}$. For $p \in (0,\infty)$, we then have:*

i) If ψ is convex and $\lim_{|r|\to\infty} \psi(r) = \infty$, then L is of lower growth type 1.
ii) If ψ is Lipschitz continuous, then L is of upper growth type 1.
iii) If ψ is convex, then for all $r > 0$ we have

$$|\psi_{|[-r,r]}|_1 \le \frac{2}{r} \, \|\psi_{|[-2r,2r]}\|_\infty \le 4|\psi_{|[-2r,2r]}|_1 \,.$$

iv) If L is convex and of upper growth type 1, then it is Lipschitz continuous.
v) If L is of upper growth type p, then there exists a constant $c_{L,p} > 0$ independent of Q such that

$$\mathcal{C}_{L,\mathrm{Q}}(t) \;\le\; c_{L,p}\bigl(|\mathrm{Q}|_p^p + |t|^p + 1\bigr)\,, \qquad t \in \mathbb{R}. \tag{2.18}$$

Moreover, L is a Nemitski loss of order p.
vi) If L is of lower growth type p, then there exists a constant $c_{L,p} > 0$ independent of Q such that

$$|\mathrm{Q}|_p^p \;\le\; c_{L,p}\bigl(\mathcal{C}_{L,\mathrm{Q}}(t) + |t|^p + 1\bigr)\,, \qquad t \in \mathbb{R}, \tag{2.19}$$

and

$$|t|^p \;\le\; c_{L,p}\bigl(\mathcal{C}_{L,\mathrm{Q}}(t) + |\mathrm{Q}|_p^p + 1\bigr)\,, \qquad t \in \mathbb{R}. \tag{2.20}$$

vii) If L is of growth type p, then we have $\mathcal{C}^*_{L,\mathrm{Q}} < \infty$ if and only if $|\mathrm{Q}|_p < \infty$.

Proof. iii). Follows immediately from Lemma A.6.5.

iv). Follows from the left inequality of *iii)* and Lemma 2.33.

ii). Follows from $|\psi(s)| = |\psi(s) - \psi(0)| \le |\psi|_1 \, |s|$ for all $s \in \mathbb{R}$.

i). The assumption $\lim_{|r|\to\infty} \psi(r) = \infty$ implies that $|\psi_{|[-r,r]}|_1 > 0$ for all sufficiently large $r > 0$. Moreover, it shows that ψ is decreasing on $(-\infty, 0]$ and increasing on $[0,\infty)$. Consequently, we have $\|\psi_{|[-r,0]}\|_\infty = \psi(r)$ for $r \le 0$, and $\|\psi_{|[0,r]}\|_\infty = \psi(r)$ for $r \ge 0$. Now, the assertion follows from applying the first part of Lemma A.6.5 to the convex functions $\mathbf{1}_{(-\infty,0]}\psi$ and $\mathbf{1}_{[0,\infty)}\psi$.

v). Writing $c_p := \max\{1, 2^{p-1}\}$, the second assertion follows from

$$L(y,t) = \psi(y-t) \le c\,\bigl(c_p|y|^p + c_p|t|^p + 1\bigr)\,, \qquad y,t \in \mathbb{R}. \tag{2.21}$$

Using this inequality, we then immediately obtain

$$\mathcal{C}_{L,\mathrm{Q}}(t) = \int_{\mathbb{R}} \psi(y-t)\, d\mathrm{Q}(y) \le c\, c_p\bigl(|\mathrm{Q}|_p^p + |t|^p\bigr) + c\,.$$

vi). We fix a $t \in \mathbb{R}$ and write $c_p := \max\{1, 2^{p-1}\}$. Since without loss of generality we may assume $\mathcal{C}_{L,\mathrm{Q}}(t) < \infty$, we can estimate

$$\begin{aligned} |\mathrm{Q}|_p^p = \int_{\mathbb{R}} |y|^p \, d\mathrm{Q}(y) &\leq c_p \int_{\mathbb{R}} |y-t|^p + |t|^p \, d\mathrm{Q}(y) \\ &= \frac{c_p}{c} \int_{\mathbb{R}} c\left(|y-t|^p - 1\right) d\mathrm{Q}(y) + c_p + c_p|t|^p \\ &\leq \frac{c_p}{c}\, \mathcal{C}_{L,\mathrm{Q}}(t) + c_p + c_p|t|^p \,. \end{aligned}$$

Now we easily find (2.19). Moreover, (2.20) can be shown analogously.

vii). The assertion immediately follows from *v)* and *vi).* □

So far we have analyzed the interplay between the growth behavior of L and the tail behavior of the conditional distributions $\mathrm{P}(\,\cdot\,|x)$. Our next step is to investigate the effect of the integration with respect to P_X. To this end, we need the following definition.

Definition 2.37. *For a distribution* P *on* $X \times \mathbb{R}$*, the* ***average*** p***-th moment****,* $p \in (0, \infty)$*, is defined by*

$$|\mathrm{P}|_p := \left(\int_X \int_{\mathbb{R}} |y|^p \, d\mathrm{P}(x,y) \right)^{1/p} = \left(\int_X \left|\mathrm{P}(\,\cdot\,|x)\right|_p^p \, d\mathrm{P}_X(x) \right)^{1/p} .$$

Moreover, its ***average*** 0***-moment*** *is defined by* $|\mathrm{P}|_0 := 1$ *and its* ***average*** ∞***-moment*** *is defined by* $|\mathrm{P}|_\infty := \operatorname{ess-sup}_{x \in X} \left|\mathrm{P}(\,\cdot\,|x)\right|_\infty$.

Again, the p-th moment of a distribution P on $X \times \mathbb{R}$ is not necessarily finite. In particular, it is easy to see that $|\mathrm{P}|_\infty < \infty$ if and only if there is an $M > 0$ such that $\operatorname{supp} \mathrm{P}(\,\cdot\,|x) \subset [-M, M]$ for P_X-almost all $x \in X$. Finally, for $0 < p \leq q \leq \infty$ we again have $|\mathrm{P}|_p \leq |\mathrm{P}|_q$.

Let us now investigate how average moments and risks interplay.

Lemma 2.38 (Average moments and risks). *Let* L *be a distance-based loss and* P *be a distribution on* $X \times Y$*. For* $p > 0$*, we then have:*

i) If L *is of upper growth type* p*, there exists a constant* $c_{L,p} > 0$ *independent of* P *such that, for all measurable* $f : X \to \mathbb{R}$*, we have*

$$\mathcal{R}_{L,\mathrm{P}}(f) \;\leq\; c_{L,p}\big(|\mathrm{P}|_p^p + \|f\|_{L_p(\mathrm{P}_X)}^p + 1\big) \,. \tag{2.22}$$

Moreover, if $|\mathrm{P}|_p < \infty$*, then* L *is a* P*-integrable Nemitski loss of order* p*, and* $\mathcal{R}_{L,\mathrm{P}}(\,\cdot\,) : L_p(\mathrm{P}_X) \to [0, \infty)$ *is well-defined and continuous.*

ii) If L *is convex and of upper growth type* p *with* $p \geq 1$*, then for all* $q \in [p-1, \infty]$ *with* $q > 0$ *there exists a constant* $c_{L,p,q} > 0$ *independent of* P *such that, for all measurable* $f : X \to \mathbb{R}$ *and* $g : X \to \mathbb{R}$*, we have*

$$\begin{aligned} &\left|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(g)\right| \\ &\leq c_{L,p,q}\Big(|\mathrm{P}|_q^{p-1} + \|f\|_{L_q(\mathrm{P}_X)}^{p-1} + \|g\|_{L_q(\mathrm{P}_X)}^{p-1} + 1\Big) \|f - g\|_{L_{\frac{q}{q-p+1}}(\mathrm{P}_X)} . \end{aligned} \tag{2.23}$$

iii) If L is of lower growth type p, there exists a constant $c_{L,p} > 0$ independent of P such that, for all measurable $f : X \to \mathbb{R}$, we have

$$|\mathrm{P}|_p^p \;\le\; c_{L,p}\big(\mathcal{R}_{L,\mathrm{P}}(f) + \|f\|_{L_p(\mathrm{P}_X)}^p + 1\big) \tag{2.24}$$

and

$$\|f\|_{L_p(\mathrm{P}_X)}^p \;\le\; c_{L,p}\big(\mathcal{R}_{L,\mathrm{P}}(f) + |\mathrm{P}|_p^p + 1\big)\,. \tag{2.25}$$

Proof. i). Inequality (2.22) follows from integrating (2.18). The second assertion follows from Inequality (2.21) and the last assertion is a consequence of Lemma 2.17.

ii). We define $r(x,y) := |y| + |f(x)| + |g(x)| + 1$, $(x,y) \in X \times Y$. By Lemma 2.36, we then obtain

$$\begin{aligned}
\big|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(g)\big| &\le \int_{X\times Y} \Big|\psi\big(y - f(x)\big) - \psi\big(y - g(x)\big)\Big| d\mathrm{P}(x,y) \\
&\le \int_{X\times Y} \big|\psi_{|[-r(x,y),r(x,y)]}\big|_1 \, \big|f(x) - g(x)\big| d\mathrm{P}(x,y) \\
&\le 2\int_{X\times Y} \frac{\|\psi_{|[-2r(x,y),2r(x,y)]}\|_\infty}{r(x,y)} \, \big|f(x) - g(x)\big| d\mathrm{P}(x,y) \\
&\le c\int_{X\times Y} \frac{|2r(x,y)|^p + 1}{2r(x,y)} \, \big|f(x) - g(x)\big| d\mathrm{P}(x,y)
\end{aligned}$$

for a suitable constant $c > 0$ only depending on L. Using $\frac{t^p+1}{t} \le 2t^{p-1}$ for all $t \ge 1$ and Hölder's inequality, we then conclude

$$\begin{aligned}
\big|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(g)\big| &\le 2^p c \int_{X\times Y} |r(x,y)|^{p-1} \, \big|f(x) - g(x)\big| d\mathrm{P}(x,y) \\
&\le 2^p c \left(\int_{X\times Y} |r|^{(p-1)s} d\mathrm{P}\right)^{1/s} \left(\int_{X\times Y} \big|f - g\big|^{s'} d\mathrm{P}\right)^{1/s'},
\end{aligned}$$

where $s := \frac{q}{p-1}$ and $\frac{1}{s'} := 1 - \frac{1}{s} = 1 - \frac{p-1}{q} = \frac{q-p+1}{q}$. Using the definition of r, we further find

$$\begin{aligned}
\left(\int_{X\times Y} |r|^{(p-1)s} d\mathrm{P}\right)^{1/s} &= \left(\int_{X\times Y} \big(|y| + |f(x)| + |g(x)| + 1\big)^q d\mathrm{P}(x,y)\right)^{(p-1)/q} \\
&\le c_q \Big(|\mathrm{P}|_q + \|f\|_{L_q(\mathrm{P}_X)} + \|g\|_{L_q(\mathrm{P}_X)} + 1\Big)^{p-1}
\end{aligned}$$

for a suitable constant $c_q > 0$. By combining the estimates, we then obtain the assertion.

iii). The inequalities (2.24) and (2.25) follow from integrating (2.19) and (2.20), respectively. □

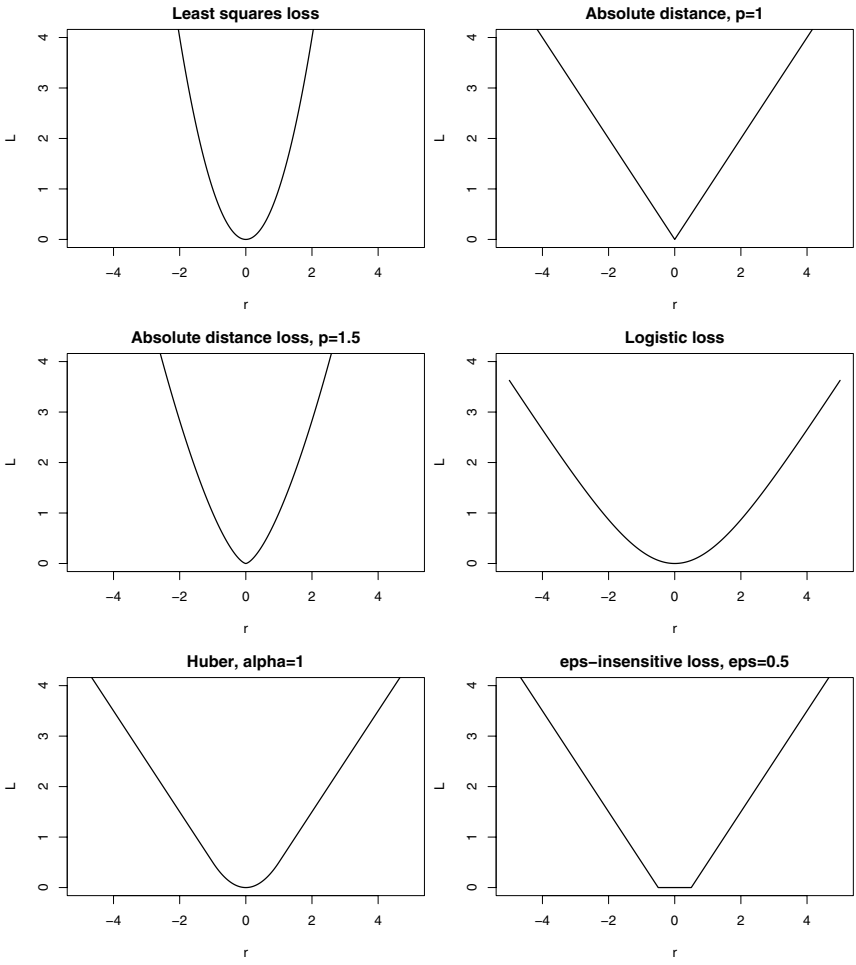

Fig. 2.2. The shape of the representing function ψ for some distance-based loss functions considered in the text.

If L is a distance-based loss function of growth type p and P is a distribution on $X \times \mathbb{R}$ with $|\mathrm{P}|_p = \infty$, the preceding lemma shows $\mathcal{R}_{L,\mathrm{P}}(f) = \infty$ for all $f \in L_p(\mathrm{P}_X)$. This suggests that we may even have $\mathcal{R}^*_{L,\mathrm{P}} = \infty$. However, in general, this is *not* the case, as Exercise 2.6 shows.

Let us finally consider some examples of distance-based loss functions (see also Figure 2.2 for some illustrations) together with some of their basic properties. We will see later in Section 3.7 that the first three losses can be used to estimate the conditional mean whenever $\mathrm{P}(\,\cdot\,|x)$ is symmetric.

Example 2.39. For $p > 0$, the **p-th power absolute distance loss** $L_{p\text{-dist}}$ is the distance-based loss function represented by

$$\psi(r) := |r|^p\,, \qquad r \in \mathbb{R}.$$

Note that for $p = 2$ this definition recovers the least squares loss. Moreover, for $p = 1$, we call $L_{p\text{-dist}}$ simply the **absolute distance loss**. It is not hard to see that $L_{p\text{-dist}}$ is of growth type p and that $L_{p\text{-dist}}$ is convex if and only if $p \geq 1$. Furthermore, $L_{p\text{-dist}}$ is strictly convex if and only if $p > 1$, and it is Lipschitz continuous if and only if $p = 1$. ◁

Example 2.40. The distance-based **logistic loss for regression** $L_{\text{r-logist}}$ is represented by

$$\psi(r) := -\ln \frac{4e^r}{(1+e^r)^2}\,, \qquad r \in \mathbb{R}.$$

Some simple calculations show that $L_{\text{r-logist}}$ is strictly convex and Lipschitz continuous, and consequently $L_{\text{r-logist}}$ is of growth type 1. ◁

Example 2.41. For $\alpha > 0$, **Huber's loss** $L_{\alpha\text{-Huber}}$ is the distance-based loss represented by

$$\psi(r) := \begin{cases} \frac{r^2}{2} & \text{if } |r| \leq \alpha \\ \alpha|r| - \frac{\alpha^2}{2} & \text{otherwise.} \end{cases}$$

Note that, for small r, Huber's loss has the shape of the least squares loss, whereas for large r it has the shape of the absolute distance loss. Consequently, $L_{\alpha\text{-Huber}}$ is convex but not strictly convex. Furthermore, it is Lipschitz continuous, and thus $L_{\alpha\text{-Huber}}$ is of growth type 1. Finally, note that the derivative of ψ equals the clipping operation (2.14) for $M = \alpha$. ◁

Example 2.42. For $\epsilon > 0$, the distance-based **ϵ-insensitive loss** $L_{\epsilon\text{-insens}}$ is represented by

$$\psi(r) := \max\bigl\{0, |r| - \epsilon\bigr\}\,, \qquad r \in \mathbb{R}.$$

The ϵ-insensitive loss ignores deviances smaller than ϵ, whereas it linearly penalizes larger deviances. It is easy to see that $L_{\epsilon\text{-insens}}$ is Lipschitz continuous and convex but not strictly convex. Therefore it is of growth type 1. We will see in Section 9.5 that this loss function can be used to estimate the conditional median, i.e., the median of $\mathrm{P}(\,\cdot\,|x)$, $x \in X$, whenever these conditional distributions are symmetric and have, for example, a Lebesgue density that is bounded away from zero on the support of $\mathrm{P}(\,\cdot\,|x)$. ◁

Example 2.43. For $\tau \in (0,1)$, the distance-based **pinball loss** $L_{\tau\text{-pin}}$ is represented by

$$\psi(r) = \begin{cases} -(1-\tau)r, & \text{if } r < 0 \\ \tau r, & \text{if } r \geq 0. \end{cases} \tag{2.26}$$

Obviously, this loss function is convex and Lipschitz continuous, but for $\tau \neq 1/2$ it is *not* symmetric. We will see in Sections 3.9 and 9.3 that this loss function can be used to estimate conditional τ-quantiles defined by

$$f^*_{\tau,\mathrm{P}}(x) := \left\{ t^* \in \mathbb{R} : \mathrm{P}\big((-\infty, t^*] \,|x\big) \geq \tau \text{ and } \mathrm{P}\big([t^*, \infty) \,|x\big) \geq 1 - \tau \right\}. \qquad \lhd$$

2.5 Further Reading and Advanced Topics

Loss functions and their associated risks have a long history in mathematical statistics and machine learning. For example, the least squares loss for regression was already used by Legendre, Gauss, and Adrain in the early 19th century (see, e.g., Harter, 1983; Stigler, 1981; and the references therein), and the classification loss function dates back to the beginning of machine learning.

In the statistical literature, density level detection has been studied by Hartigan (1987), Müller and Sawitzki (1991), Polonik (1995), Sawitzki (1996), and Tsybakov (1997), among others. Most of these authors focus on the so-called *excess mass approach*. Steinwart *et al.* (2005) showed that this approach is equivalent to an empirical risk minimization approach using a particular classification problem, and based on this observation the authors derived an SVM for the density level detection problem (see also Sections 3.8 and 8.6). Moreover, the risk based on the density level detection loss defined in (2.6) was proposed by Polonik (1995) and later also used by, e.g., Tsybakov (1997) and Ben-David and Lindenbaum (1997). Various applications of the DLD problem, such as cluster analysis, testing for multimodality, and spectral analysis, are described by Hartigan (1975), Müller and Sawitzki (1991), and Polonik (1995). Finally, using DLD for anomaly detection is a widely known technique; see Davies and Gather (1993) and Ripley (1996), for example.

It is well-known that empirical risk minimization for the classification loss typically leads to combinatorial optimization problems that in many cases are NP-hard to solve (see, e.g., Höffgen *et al.*, 1995). Using a margin-based loss as a surrogate for the classification loss is a well-known trick in machine learning to make the training process algorithmically more tractable (see, e.g., the motivation for the hinge loss by Cortes and Vapnik, 1995). In particular, for SVMs, the first surrogates for the classification loss were the hinge loss and its squared variant, the truncated least squares loss. Later, other loss functions, such as the least squares loss and the logistic loss, were introduced into the support vector machine literature by Suykens and Vandewalle (1999), see also Poggio and Girosi (1990), Wahba (1990), and Girosi *et al.* (1995) for earlier work in this direction, and Wahba (1999), respectively. Other margin-based loss functions used in the literature include the exponential loss $\varphi(t) := \exp(-t)$, $t \in \mathbb{R}$, used in the AdaBoost algorithm (see Freund and Schapire, 1996; Breiman, 1999b) and the loss $\varphi(t) := (1-t)^5$

used in the ARC-X4 procedure of Breiman (1998). Some further margin-based losses used in boosting algorithms are listed by Mason *et al.* (2000). Finally, Zhang's inequality was shown by Zhang (2004b).

The importance of Nemitski losses for conditional distributions with unbounded support was first discovered by De Vito *et al.* (2004), and the growth type of distance-based losses was introduced by Christmann and Steinwart (2007).

Huber's loss was proposed by Huber (1964) in the context of robust statistics, and the logistic loss function was already used in the Princeton study by Andrews *et al.* (1972). Moreover, the pinball loss was utilized by Koenker and Bassett (1978) in the context of quantile regression. Last but not least, for a comparison between the absolute distance loss and the least squares loss regarding computational speed for certain algorithms, we refer to Portnoy and Koenker (1997).

2.6 Summary

In this chapter, we introduced loss functions and their associated risks. We saw in Section 2.1 that loss functions can be used to formalize many learning goals, including classification, regression, and density level detection problems. We then investigated simple yet important properties of loss functions. Among them, the notion of integrable Nemitski losses will be a central tool in the following chapters.

Since the classification loss typically leads to computationally hard optimization problems, we presented margin-based surrogates in Section 2.3. For one of these surrogates, namely the hinge loss, we explicitly showed in Zhang's inequality how its excess risk relates to the excess classification risk. In Chapter 3, we will see that a similar relation holds for the other margin-based losses we presented.

Finally, we investigated distance-based loss functions for regression problems in Section 2.4. There, we first showed how the growth behavior of the loss function L together with the average conditional tail behavior of the distribution P determines whether L is a P-integrable Nemitski loss. These considerations will play a crucial role in Chapter 9, where we investigate the learning capabilities of SVMs in regression problems. At the end of Section 2.4, we presented some examples of distance-based losses, including the least squares loss, the pinball loss, the logistic loss, Huber's loss, and the ϵ-insensitive loss. In Chapter 3, we will investigate their relationships to each other.

2.7 Exercises

2.1. Convex and Lipschitz continuous risks ($\star$)
Prove Lemma 2.13 and Lemma 2.19.

2.2. Properties of some margin-based losses ($\star$)
Verify the assertions made in the examples of Section 2.3. Moreover, investigate the properties of the *exponential loss* represented by $\varphi(t) := \exp(-t)$, $t \in \mathbb{R}$, and the *sigmoid loss* represented by $\varphi(t) := 1 - \tanh(t)$, $t \in \mathbb{R}$.

2.3. A surrogate inequality for the logistic loss ($\star\star\star\star$)
Try to find an inequality between the excess classification risk and the excess $L_{\text{c-logist}}$-risk. Compare your findings with the inequality we will obtain in Section 3.4.

2.4. Properties of some distance-based losses ($\star$)
Verify the assertions made in the examples of Section 2.4.

2.5. Clippable convex distance-based losses ($\star\star$)
Let L be a distance-based loss function whose representing function ψ satisfies $\lim_{r\to\pm\infty} \psi(r) = \infty$. Show that L can be clipped at some $M > 0$ if and only if Y is bounded.

2.6. Infinite Bayes risk for regression ($\star\star$)
Let L be a distance-based loss of growth type p and $X := Y := \mathbb{R}$. Find a distribution P on $X \times Y$ such that $|\mathrm{P}|_p = \infty$ and $\mathcal{R}_{L,\mathrm{P}}(f) = \infty$ for all $f \in L_p(\mathrm{P}_X)$ but $\mathcal{R}^*_{L,\mathrm{P}} < \infty$.

Hint: Use a measurable function $g : X \to \mathbb{R}$ with $g \notin L_p(\mathrm{P}_X)$.