3 Risk and Loss Functions

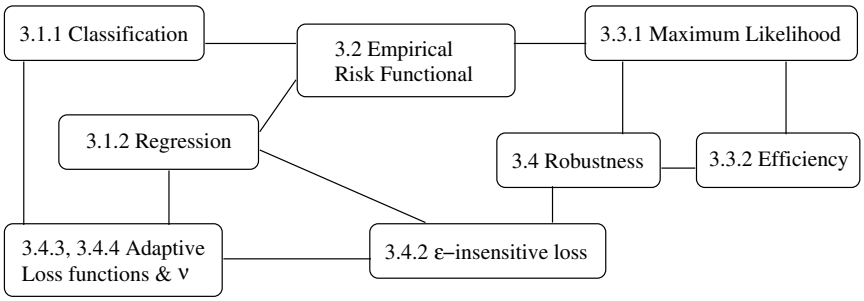
One of the most immediate requirements in any learning problem is to specify what exactly we would like to achieve, minimize, bound, or approximate. In other words, we need to determine a criterion according to which we will assess the quality of an estimate $f : \mathcal{X} \to \mathcal{Y}$ obtained from data.

This question is far from trivial. Even in binary classification there exist ample choices. The selection criterion may be the fraction of patterns classified correctly, it could involve the confidence with which the classification is carried out, or it might take into account the fact that losses are not symmetric for the two classes, such as in health diagnosis problems. Furthermore, the loss for an error may be input-dependent (for instance, meteorological predictions may require a higher accuracy in urban regions), and finally, we might want to obtain probabilities rather than a binary prediction of the class labels -1 and 1. Multi class discrimination and regression add even further levels of complexity to the problem. Thus we need a means of encoding these criteria.

Overview

The chapter is structured as follows: in Section 3.1, we begin with a brief overview of common loss functions used in classification and regression algorithms. This is done without much mathematical rigor or statistical justification, in order to provide basic working knowledge for readers who want to get a quick idea of the default design choices in the area of kernel machines. Following this, Section 3.2 formalizes the idea of risk. The risk approach is the predominant technique used in this book, and most of the algorithms presented subsequently minimize some form of a risk functional. Section 3.3 treats the concept of loss functions from a statistical perspective, points out the connection to the estimation of densities and introduces the notion of efficiency. Readers interested in more detail should also consider Chapter 16, which discusses the problem of estimation from a Bayesian perspective. The later parts of this section are intended for readers interested in the more theoretical details of estimation. The concept of robustness is introduced in Section 3.4. Several commonly used loss functions, such as Huber's loss and the ε-insensitive loss, enjoy robustness properties with respect to rather general classes of distributions. Beyond the basic relations, will show how to adjust the ε-insensitive loss in such a way as to accommodate different amounts of variance automatically. This will later lead to the construction of so-called ν Support Vector Algorithms (see Chapters 7, 8, and 9).

While technical details and proofs can be omitted for most of the present chapter, we encourage the reader to review the practical implications of this section.



Prerequisites

As usual, exercises for all sections can be found at the end. The chapter requires knowledge of probability theory, as introduced in Section B.1.

3.1 Loss Functions

Let us begin with a formal definition of what we mean by the loss incurred by a function f at location x, given an observation y.

Definition 3.1 (Loss Function) *Denote by $(x, y, f(x)) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ the triplet consisting of a pattern x, an observation y and a prediction $f(x)$. Then the map $c : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ with the property $c(x, y, y) = 0$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ will be called a loss function.*

Note that we require c to be a nonnegative function. This means that we will never get a payoff from an extra good prediction. If the latter was the case, we could always recover non-negativity (provided the loss is bounded from below), by using a simple shift operation (possibly depending on x). Likewise we can always satisfy the condition that exact predictions ($f(x) = y$) never cause any loss. The advantage of these extra conditions on c is that we know that the minimum of the loss is 0 and that it is obtainable, at least for a given x, y.

Minimized Loss ≠ Incurred Loss

Next we will formalize different kinds of *loss*, as described informally in the introduction of the chapter. Note that the incurred loss is not always the quantity that we will attempt to minimize. For instance, for algorithmic reasons, some loss functions will prove to be infeasible (the binary loss, for instance, can lead to NP-hard optimization problems [367]). Furthermore, statistical considerations such as the desire to obtain confidence levels on the prediction (Section 3.3.1) will also influence our choice.

3.1.1 Binary Classification

Misclassification Error

The simplest case to consider involves counting the misclassification error if pattern x is classified wrongly we incur loss 1, otherwise there is no penalty.:

$$c(x, y, f(x)) = \begin{cases} 0 & \text{if } y = f(x) \\ 1 & \text{otherwise} \end{cases} \tag{3.1}$$

This definition of c does not distinguish between different classes and types of errors (false positive or negative).[1]

Asymmetric and Input-Dependent Loss

A slight extension takes the latter into account. For the sake of simplicity let us assume, as in (3.1), that we have a binary classification problem. This time, however, the loss may depend on a function $\tilde{c}(x)$ which accounts for input-dependence, i.e.

$$c(x, y, f(x)) = \begin{cases} 0 & \text{if } y = f(x) \\ \tilde{c}(x) & \text{otherwise} \end{cases} \tag{3.2}$$

A simple (albeit slightly contrived) example is the classification of objects into rocks and diamonds. Clearly, the incurred loss will depend largely on the weight of the object under consideration.

Analogously, we might distinguish between errors for $y = 1$ and $y = -1$ (see, e.g., [331] for details). For instance, in a fraud detection application, we would like to be really sure about the situation before taking any measures, rather than losing potential customers. On the other hand, a blood bank should consider even the slightest suspicion of disease before accepting a donor.

Confidence Level

Rather than predicting only whether a given object x belongs to a certain class y, we may also want to take a certain confidence level into account. In this case, $f(x)$ becomes a real-valued function, even though $y \in \{-1, 1\}$.

In this case, $\operatorname{sgn}(f(x))$ denotes the class label, and the absolute value $|f(x)|$ the confidence of the prediction. Corresponding loss functions will depend on the product $yf(x)$ to assess the quality of the estimate. The *soft margin* loss function, as introduced by Bennett and Mangasarian [40, 111], is defined as

Soft Margin Loss

$$c(x, y, f(x)) = \max(0, 1 - yf(x)) = \begin{cases} 0 & \text{if } yf(x) \geq 1, \\ 1 - yf(x) & \text{otherwise.} \end{cases} . \tag{3.3}$$

In some cases [348, 125] (see also Section 10.6.2) the squared version of (3.3) provides an expression that can be minimized more easily;

$$c(x, y, f(x)) = \max(0, 1 - yf(x))^2. \tag{3.4}$$

Logistic Loss

The soft margin loss closely resembles the so-called *logistic* loss function (cf. [251], as well as Problem 3.1 and Section 16.1.1);

$$c(x, y, f(x)) = \ln\left(1 + \exp\left(-yf(x)\right)\right). \tag{3.5}$$

We will derive this loss function in Section 3.3.1. It is used in order to associate a probabilistic meaning with $f(x)$.

Note that in both (3.3) and (3.5) (nearly) no penalty occurs if $yf(x)$ is sufficiently large, i.e. if the patterns are classified correctly with large confidence. In particular, in (3.3) a minimum confidence of 1 is required for zero loss. These loss functions

1. A *false positive* is a point which the classifier erroneously assigns to class 1, a *false negative* is erroneously assigned to class -1.

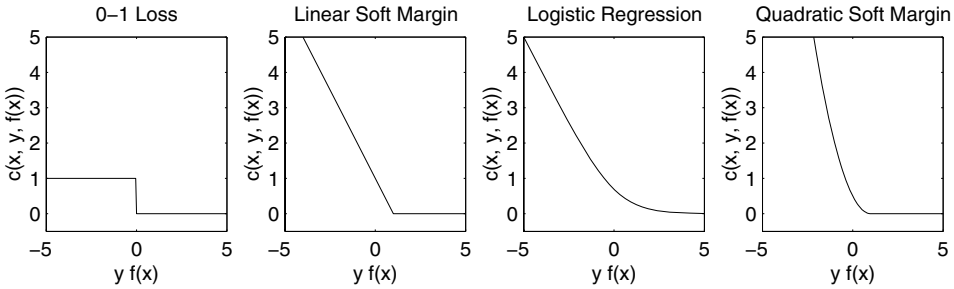


Figure 3.1 From left to right: 0-1 loss, linear soft margin loss, logistic regression, and quadratic soft margin loss. Note that both soft margin loss functions are upper bounds on the 0-1 loss.

led to the development of *large margin classifiers* (see [491, 460, 504] and Chapter 5 for further details). Figure 3.1 depicts various popular loss functions.[2]

Multi Class Discrimination

Matters are more complex when dealing with more than two classes. Each type of misclassification could potentially incur a different loss, leading to an $M \times M$ matrix (M being the number of classes) with positive off-diagonal and zero diagonal entries. It is still a matter of ongoing research in which way a confidence level should be included in such cases (cf. [41, 311, 593, 161, 119]).

3.1.2 Regression

When estimating real-valued quantities, it is usually the size of the difference $y - f(x)$, i.e. the amount of misprediction, rather than the product $yf(x)$, which is used to determine the quality of the estimate. For instance, this can be the actual loss incurred by mispredictions (e.g., the loss incurred by mispredicting the value of a financial instrument at the stock exchange), provided the latter is known and computationally tractable.[3] Assuming location independence, in most cases the loss function will be of the type

$$c(x, y, f(x)) = \tilde{c}(f(x) - y). \tag{3.7}$$

See Figure 3.2 below for several regression loss functions. Below we list the ones most common in kernel methods.

---

2. Other popular loss functions from the generalized linear model context include the inverse complementary log-log function. It is given by

$$c(x, y, f(x)) = 1 - \exp(-\exp(yf(x))). \tag{3.6}$$

This function, unfortunately, is not convex and therefore it will not lead to a convex optimization problem. However, it has nice robustness properties and therefore we think that it should be investigated in the present context.

3. As with classification, computational tractability is one of the primary concerns. This is not always satisfying from a statistician's point of view, yet it is crucial for any practical implementation of an estimation algorithm.

Squared Loss

The popular choice is to minimize the sum of squares of the residuals $f(x) - y$. As we shall see in Section 3.3.1, this corresponds to the assumption that we have additive normal noise corrupting the observations y_i. Consequently we minimize

$$c(x, y, f(x)) = (f(x) - y)^2 \text{ or equivalently } \tilde{c}(\xi) = \xi^2. \tag{3.8}$$

For convenience of subsequent notation, $\frac{1}{2}\xi^2$ rather than ξ^2 is often used.

ε-insensitive Loss and ℓ_1 Loss

An extension of the soft margin loss (3.3) to regression is the *ε-insensitive* loss function [561, 572, 562]. It is obtained by symmetrization of the "hinge" of (3.3),

$$\tilde{c}(\xi) = \max(|\xi| - \varepsilon, 0) =: |\xi|_\varepsilon. \tag{3.9}$$

The idea behind (3.9) is that deviations up to ε should not be penalized, and all further deviations should incur only a linear penalty. Setting $\varepsilon = 0$ leads to an ℓ_1 loss, i.e., to minimization of the sum of absolute deviations. This is written

$$\tilde{c}(\xi) = |\xi|. \tag{3.10}$$

We will study these functions in more detail in Section 3.4.2.

Practical Considerations

For efficient implementations of learning procedures, it is crucial that loss functions satisfy certain properties. In particular, they should be cheap to compute, have a small number of discontinuities (if any) in the first derivative, and be convex in order to ensure the uniqueness of the solution (see Chapter 6 and also Problem 3.6 for details). Moreover, we may want to obtain solutions that are computationally efficient, which may disregard a certain number of training points. This leads to conditions such as vanishing derivatives for a range of function values $f(x)$. Finally, requirements such as outlier resistance are also important for the construction of estimators.

3.2 Test Error and Expected Risk

Now that we have determined how errors should be penalized on specific instances $(x, y, f(x))$, we have to find a method to combine these (local) penalties. This will help us to assess a particular estimate f.

In the following, we will assume that there exists a probability distribution $\mathrm{P}(x, y)$ on $\mathcal{X} \times \mathcal{Y}$ which governs the data generation and underlying functional dependency. Moreover, we denote by $\mathrm{P}(y|x)$ the *conditional* distribution of y given x, and by $d\mathrm{P}(x, y)$ and $d\mathrm{P}(y|x)$ the integrals with respect to the distributions $\mathrm{P}(x, y)$ and $\mathrm{P}(y|x)$ respectively (cf. Section B.1.3).

3.2.1 Exact Quantities

Unless stated otherwise, we assume that the data (x, y) are drawn iid (independent and identically distributed, see Section B.1) from $\mathrm{P}(x, y)$. Whether or not we have

knowledge of the test patterns at training time[4] makes a significant difference in the design of learning algorithms. In the latter case, we will want to minimize the *test error* on that *specific* test set; in the former case, the *expected* error over *all possible* test sets.

Definition 3.2 (Test Error) *Assume that we are not only given the* training *data* $\{x_1, \ldots, x_m\}$ *along with target values* $\{y_1, \ldots y_m\}$ *but also the* test *patterns* $\{x'_1, \ldots x'_{m'}\}$ *on which we would like to predict* y'_i $(i = 1, \ldots, m')$*. Since we already know* x'_i*, all we should care about is to minimize the* expected error *on the* test set*. We formalize this in the following definition*

Transduction Problem

$$R_{\text{test}}[f] := \frac{1}{m'} \sum_{i=1}^{m'} \int_{\mathcal{Y}} c(x'_i, y, f(x'_i)) d\mathrm{P}(y|x'_i). \tag{3.11}$$

Unfortunately, this problem, referred to as *transduction*, is quite difficult to address, both computationally and conceptually, see [562, 267, 37, 211]. Instead, one typically considers the case where no knowledge about test patterns is available, as described in the following definition.

Definition 3.3 (Expected Risk) *If we have no knowledge about the test patterns (or decide to ignore them) we should minimize the expected error over all possible training patterns. Hence we have to minimize the expected loss with respect to* P *and c*

$$R[f] := \mathbf{E}\left[R_{\text{test}}[f]\right] = \mathbf{E}\left[c(x, y, f(x))\right] = \int_{\mathcal{X}\times\mathcal{Y}} c(x, y, f(x)) d\mathrm{P}(x, y). \tag{3.12}$$

Here the integration is carried out with respect to the distribution $\mathrm{P}(x, y)$. Again, just as (3.11), this problem is intractable, since we do not know $\mathrm{P}(x, y)$ explicitly. Instead, we are only given the training patterns (x_i, y_i). The latter, however, allow us to replace the unknown distribution $\mathrm{P}(x, y)$ by its empirical estimate.

To study connections between loss functions and density models, it will be convenient to assume that there exists a density $p(x, y)$ corresponding to $\mathrm{P}(x, y)$. This means that we may replace $\int d\mathrm{P}(x, y)$ by $\int p(x, y) dx dy$ and the appropriate measure on $\mathcal{X} \times \mathcal{Y}$. Such a density $p(x, y)$ need not always exist (see Section B.1 for more details) but we will not give further heed to these concerns at present.

3.2.2 Approximations

Unfortunately, this change in notation did not solve the problem. All we have at our disposal is the actual training data. What one usually does is replace $p(x, y)$ by the *empirical density*

Empirical Density

$$p_{\text{emp}}(x, y) := \frac{1}{m} \sum_{i=1}^{m} \delta_{x_i}(x) \delta_{y_i}(y). \tag{3.13}$$

4. The test *outputs*, however, are not available during training.

Here $\delta_{x'}(x)$ denotes the δ-distribution, satisfying $\int \delta_{x'}(x) f(x) dx = f(x')$. The hope is that replacing p by p_{emp} will lead to a quantity that is "reasonably close" to the expected risk. This will be the case if the class of possible solutions f is sufficiently limited [568, 571]. The issue of closeness with regard to different estimators will be discussed in further detail in Chapters 5 and 12. Substituting $p_{\text{emp}}(x, y)$ into (3.12) leads to the empirical risk:

Definition 3.4 (Empirical Risk) *The empirical risk is defined as*

$$R_{\text{emp}}[f] := \int_{\mathcal{X}\times\mathcal{Y}} c(x, y, f(x)) p_{\text{emp}}(x, y) dx dy = \frac{1}{m} \sum_{i=1}^{m} c(x_i, y_i, f(x_i)). \tag{3.14}$$

M-Estimator

This quantity has the advantage that, given the training data, we can readily compute and also minimize it. This constitutes a particular case of what is called an *M-estimator* in statistics. Estimators of this type are studied in detail in the field of empirical processes [554]. As pointed out in Section 3.1, it is crucial to understand that although our particular M-estimator is built from minimizing a loss, this need not always be the case. From a decision-theoretic point of view, the question of which loss to choose is a separate issue, which is dictated by the problem at hand as well as the goal of trying to evaluate the performance of estimation methods, rather than by the problem of trying to define a particular estimation method [582, 166, 43].

Ill-Posed Problems

These considerations aside, it may appear as if (3.14) is the answer to our problems, and all that remains to be done is to find a suitable class of functions $\mathcal{F} \ni f$ such that we can minimize $R_{\text{emp}}[f]$ with respect to $\mathcal{F}$. Unfortunately, determining $\mathcal{F}$ is quite difficult (see Chapters 5 and 12 for details). Moreover, the minimization of $R_{\text{emp}}[f]$ can lead to an ill-posed problem [538, 370]. We will show this with a simple example.

Example of an Ill-Posed Problem

Assume that we want to solve a regression problem using the quadratic loss function (3.8) given by $c(x, y, f(x) = (y - f(x))^2$. Moreover, assume that we are dealing with a linear class of functions,[5] say

$$\mathcal{F} := \left\{ f \,\middle|\, f(x) = \sum_{i=1}^{n} \alpha_i f_i(x) \text{ with } \alpha_i \in \mathbb{R} \right\}, \tag{3.15}$$

where the f_i are functions mapping $\mathcal{X}$ to $\mathbb{R}$.

We want to find the minimizer of R_{emp}, i.e.,

$$\underset{f\in\mathcal{F}}{\text{minimize}}\, R_{\text{emp}}[f] = \underset{\boldsymbol{\alpha}\in\mathbb{R}^n}{\text{minimize}}\, \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} \alpha_j f_j(x_i) \right)^2. \tag{3.16}$$

5. In the simplest case, assuming $\mathcal{X}$ is contained in a vector space, these could be functions that extract coordinates of x; in other words, $\mathcal{F}$ would be the class of linear functions on $\mathcal{X}$.

Computing the derivative of $R_{\text{emp}}[f]$ with respect to $\boldsymbol{\alpha}$ and defining $F_{ij} := f_i(x_j)$, we can see that the minimum of (3.16) is achieved if

$$F^\top \mathbf{y} = F^\top F \boldsymbol{\alpha}. \tag{3.17}$$

A sufficient condition for (3.17) is $\boldsymbol{\alpha} = \left(F^\top F\right)^{-1} F^\top \mathbf{y}$ where $\left(F^\top F\right)^{-1}$ denotes the (pseudo-)inverse of the matrix.

Condition of a Matrix

If $F^\top F$ has a bad condition number (i.e. the quotient between the largest and the smallest eigenvalue of $F^\top F$ is large), it is numerically difficult [423, 530] to solve (3.17) for $\boldsymbol{\alpha}$. Furthermore, if $n > m$, i.e. if we have more basis functions f_i than training patterns x_i, there will exist a subspace of solutions with dimension at least $n - m$, satisfying (3.17). This is undesirable both practically (speed of computation) and theoretically (we would have to deal with a whole class of solutions rather than a single one).

One might also expect that if $\mathcal{F}$ is too rich, the discrepancy between $R_{\text{emp}}[f]$ and $R[f]$ could be large. For instance, if F is an $m \times m$ matrix of full rank, $\mathcal{F}$ contains an f that predicts all target values y_i correctly on the training data. Nevertheless, we cannot expect that we will also obtain zero prediction error on unseen points. Chapter 4 will show how these problems can be overcome by adding a so-called regularization term to $R_{\text{emp}}[f]$.

3.3 A Statistical Perspective

Given a particular pattern $\tilde{x}$, we may want to ask what risk we can expect for it, and with which *probability* the corresponding loss is going to occur. In other words, instead of (or in addition to) $\mathbf{E}\left[c(\tilde{x}, y, f(\tilde{x})\right]$ for a fixed $\tilde{x}$, we may want to know the distribution of y given $\tilde{x}$, i.e., $\mathrm{P}(y|\tilde{x})$.

(Bayesian) statistics (see [338, 432, 49, 43] and also Chapter 16) often attempt to estimate the density corresponding to the random variables (x, y), and in some cases, we may really *need* information about $p(x, y)$ to arrive at the desired conclusions given the training data (e.g., medical diagnosis). However, we always have to keep in mind that if we model the density p first, and subsequently, based on this approximation, compute a minimizer of the expected risk, we will have to make two approximations. This could lead to inferior or at least not easily predictable results. Therefore, wherever possible, we should avoid solving a more general problem, since additional approximation steps might only make the estimates worse [561].

3.3.1 Maximum Likelihood Estimation

All this said, we still may want to compute the conditional density $p(y|x)$. For this purpose we need to model how y is generated, based on some underlying dependency $f(x)$; thus, we specify the functional form of $p(y|x, f(x))$ and maximize

the expression with respect to f. This will provide us with the function f that is *most likely* to have generated the data.

Definition 3.5 (Likelihood) *The likelihood of a sample $(x_1, y_1), \ldots (x_m, y_m)$ given an underlying functional dependency f is given by*

$$p(\{x_1, \ldots, x_m\}, \{y_1, \ldots, y_m\}|f) = \prod_{i=1}^{m} p(x_i, y_i|f) = \prod_{i=1}^{m} p(y_i|x_i, f)p(x_i) \tag{3.18}$$

Strictly speaking the likelihood only depends on the values $f(x_1), \ldots, f(x_m)$ rather than being a functional of f itself. To keep the notation simple, however, we write $p(\{x_1, \ldots, x_m\}, \{y_1, \ldots, y_m\}|f)$ instead of the more heavyweight expression $p(\{x_1, \ldots, x_m\}, \{y_1, \ldots, y_m\}|\{f(x_1), \ldots, f(x_m)\})$.

For practical reasons, we convert products into sums by taking the negative logarithm of $\mathrm{P}(\{x_1, \ldots, x_m\}, \{y_1, \ldots, y_m\}|f)$, an expression which is then conveniently minimized. Furthermore, we may drop the $p(x_i)$ from (3.18), since they do not depend on f. Thus maximization of (3.18) is equivalent to minimization of the

Log-Likelihood

Log-Likelihood

$$\mathcal{L}[f] := \sum_{i=1}^{m} -\ln p(y_i|x_i, f). \tag{3.19}$$

Regression

Remark 3.6 (Regression Loss Functions) *Minimization of $\mathcal{L}[f]$ and of $R_{\mathrm{emp}}[f]$ coincide if the loss function c is chosen according to*

$$c(x, y, f(x)) = -\ln p(y|x, f). \tag{3.20}$$

Assuming that the target values y were generated by an underlying functional dependency f plus additive noise ξ with density p_ξ, i.e. $y_i = f_{\mathrm{true}}(x_i) + \xi_i$, we obtain

$$c(x, y, f(x)) = -\ln p_\xi(y - f(x)). \tag{3.21}$$

Things are slightly different in classification. Since all we are interested in is the probability that pattern x has label 1 or -1 (assuming binary classification), we can transform the problem into one of estimating the logarithm of the probability

Classification

that a pattern assumes its correct label.

Remark 3.7 (Classification Loss Functions) *We have a finite set of labels, which allows us to model $\mathrm{P}(y|f(x))$ directly, instead of modelling a density. In the binary classification case (classes 1 and -1) this problem becomes particularly easy, since all we have to do is assume functional dependency underlying $\mathrm{P}(1|f(x))$: this immediately gives us $\mathrm{P}(-1|f(x)) = 1 - \mathrm{P}(1|f(x))$. The link to loss functions is established via*

$$c(x, y, f(x)) = -\ln \mathrm{P}(y|f(x)). \tag{3.22}$$

The same result can be obtained by minimizing the cross entropy[6] *between the classifica-*

6. In the case of discrete variables the cross entropy between two distributions P and Q is defined as $\sum_i \mathrm{P}(i) \ln \mathrm{Q}(i)$.

Table 3.1 Common loss functions and corresponding density models according to Remark 3.6. As a shorthand we use $\tilde{c}(f(x) - y) := c(x, y, f(x))$.

	loss function $\tilde{c}(\xi)$	density model $p(\xi)$
ε-insensitive	$\lvert\xi\rvert_\varepsilon$	$\frac{1}{2(1+\varepsilon)} \exp(-\lvert\xi\rvert_\varepsilon)$
Laplacian	$\lvert\xi\rvert$	$\frac{1}{2} \exp(-\lvert\xi\rvert)$
Gaussian	$\frac{1}{2}\xi^2$	$\frac{1}{\sqrt{2\pi}} \exp(-\frac{\xi^2}{2})$
Huber's robust loss	$\begin{cases} \frac{1}{2\sigma}(\xi)^2 & \text{if } \lvert\xi\rvert \le \sigma \\ \lvert\xi\rvert - \frac{\sigma}{2} & \text{otherwise} \end{cases}$	$\propto \begin{cases} \exp(-\frac{\xi^2}{2\sigma}) & \text{if } \lvert\xi\rvert \le \sigma \\ \exp(\frac{\sigma}{2} - \lvert\xi\rvert) & \text{otherwise} \end{cases}$
Polynomial	$\frac{1}{d}\lvert\xi\rvert^d$	$\frac{d}{2\Gamma(1/d)} \exp(-\lvert\xi\rvert^d)$
Piecewise polynomial	$\begin{cases} \frac{1}{d\sigma^{d-1}}\lvert\xi\rvert^d & \text{if } \lvert\xi\rvert \le \sigma \\ \lvert\xi\rvert - \sigma\frac{d-1}{d} & \text{otherwise} \end{cases}$	$\propto \begin{cases} \exp(-\frac{\lvert\xi\rvert^d}{d\sigma^{d-1}}) & \text{if } \lvert\xi\rvert\sigma \\ \exp(\sigma\frac{d-1}{d} - \lvert\xi\rvert) & \text{otherwise} \end{cases}$

tion labels y_i and the probabilities $p(y|f(x))$, as is typically done in a generalized linear models context (see e.g., [355, 232, 163]). For binary classification (with $y \in \{\pm 1\}$) we obtain

$$c(x, y, f(x)) = \frac{1+y}{2} \ln \mathrm{P}(y = 1|f(x)) + \frac{1-y}{2} \ln \mathrm{P}(y = -1|f(x)). \tag{3.23}$$

When substituting the actual values for y into (3.23), this reduces to (3.22).

At this point we have a choice in modelling $\mathrm{P}(y = 1|f(x))$ to suit our needs. Possible models include the logistic transfer function, the probit model, the inverse complementary log-log model. See Section 16.3.5 for a more detailed discussion of the choice of such *link functions*. Below we explain connections in some more detail for the logistic link function.

For a logistic model, where $\mathrm{P}(y = \pm 1|x, f) \propto \exp(\pm\frac{1}{2}f(x))$, we obtain after normalization

$$\mathrm{P}(y = 1|x, f) := \frac{\exp(f(x))}{1 + \exp(f(x))} \tag{3.24}$$

and consequently $-\ln \mathrm{P}(y = 1|x, f) = \ln(1 + \exp(-f(x)))$. We thus recover (3.5) as the loss function for classification. Choices other than (3.24) for a map $\mathbb{R} \to [0, 1]$ will lead to further loss functions for classification. See [579, 179, 596] and Section 16.1.1 for more details on this subject.

It is important to note that not every loss function used in classification corresponds to such a density model (recall that in this case, the probabilities have to add up to 1 for any value of $f(x)$). In fact, one of the most popular loss functions, the soft margin loss (3.3), does not enjoy this property. A discussion of these issues can be found in [521].

Examples

Table 3.1 summarizes common loss functions and the corresponding density models as defined by (3.21), some of which were already presented in Section 3.1. It is an exhaustive list of the loss functions that will be used in this book for regression. Figure 3.2 contains graphs of the functions.

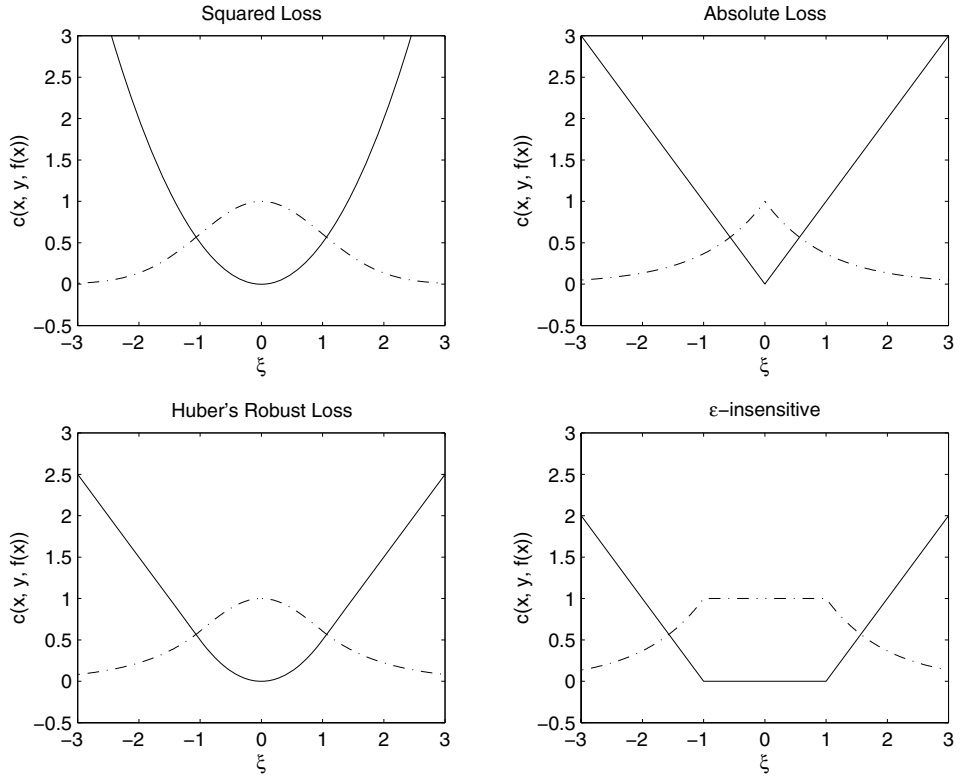


Figure 3.2 Graphs of loss functions and corresponding density models. upper left: Gaussian, upper right: Laplacian, lower left: Huber's robust, lower right: ε-insensitive.

Practical Considerations

We conclude with a few cautionary remarks. The loss function resulting from a maximum likelihood reasoning might be non-convex. This might spell trouble when we try to find an efficient solution of the corresponding minimization problem. Moreover, we made a very strong assumption by claiming to know $\mathrm{P}(y|x, f)$ explicitly, which was necessary in order to evaluate (3.20).

Finally, the solution we obtain by minimizing the log-likelihood depends on the class of functions $\mathcal{F}$. So we are in no better situation than by minimizing $R_{\mathrm{emp}}[f]$, albeit with the additional constraint, that the loss functions $c(x, y, f(x))$ must correspond to a probability density.

3.3.2 Efficiency

The above reasoning could mislead us into thinking that the choice of loss function is rather arbitrary, and that there exists no good means of assessing the performance of an estimator. In the present section we will develop tools which can be used to compare estimators that are derived from different loss functions. For this purpose we need to introduce additional statistical concepts which deal with the efficiency of an estimator. Roughly speaking, these give an indication of how

"noisy" an estimator is with respect to a reference estimator.

We begin by formalizing the concept of an estimator. Denote by $\mathrm{P}(y|\theta)$ a distribution of y depending (amongst other variables) on the parameters θ, and by $Y = \{y_1, \ldots, y_m\}$ an m-sample drawn iid from $\mathrm{P}(y|\theta)$. Note that the use of the symbol y bears no relation to the y_i that are outputs of some functional dependency (cf. Chapter 1). We employ this symbol because some of the results to be derived will later be applied to the outputs of SV regression.

Estimator

Next, we introduce the *estimator* $\hat{\theta}(Y)$ of the parameters θ, based on Y. For instance, $\mathrm{P}(y|\theta)$ could be a Gaussian with fixed variance and mean θ, and $\hat{\theta}(Y)$ could be the estimator $(1/m)\sum_{i=1}^{m} y_i$.

To avoid cumbersome notation, we use the shorthand

$$\mathbf{E}_\theta\left[\xi(y)\right] := \mathbf{E}_{\mathrm{P}(y|\theta)}\left[\xi(y)\right] = \int \xi(y) d\mathrm{P}(y|\theta), \tag{3.25}$$

to express expectations of a random variable $\xi(y)$ with respect to $\mathrm{P}(y|\theta)$. One criterion that we might impose on an estimator is that it be unbiased, i.e., that on average, it tells us the correct value of the parameter it attempts to estimate.

Definition 3.8 (Unbiased Estimator) *An unbiased estimator $\hat{\theta}(Y)$ of the parameters θ in $\mathrm{P}(y|\theta)$ satisfies*

$$\mathbf{E}_\theta\left[\hat{\theta}(Y)\right] = \theta. \tag{3.26}$$

In this section, we will focus on unbiased estimators. In general, however, the estimators we are dealing with in this book will not be unbiased. In fact, they will have a bias towards 'simple', low-complexity functions. Properties of such estimators are more difficult to deal with, which is why, for the sake of simplicity, we restrict ourselves to the unbiased case in this section. Note, however, that "biasedness" is not a bad property by itself. On the contrary, there exist cases as the one described by James and Stein [262] where biased estimators consistently outperform unbiased estimators in the finite sample size setting, both in terms of variance and prediction error.

A possible way to compare unbiased estimators is to compute their variance. Other quantities such as moments of higher order or maximum deviation properties would be valid criteria as well, yet for historical and practical reasons the variance has become a standard tool to benchmark estimators. The Fisher information matrix is crucial for this purpose since it will tell us via the Cramér-Rao bound (Theorem 3.11) the minimal possible variance for an unbiased estimator. The idea is that the smaller the variance, the lower (typically) the probability that $\hat{\theta}(Y)$ will deviate from θ by a large amount. Therefore, we can use the variance as a possible one number summary to compare different estimators.

Definition 3.9 (Score Function, Fisher Information, Covariance) *Assume there exists a density $p(y|\theta)$ for the distribution $\mathrm{P}(y|\theta)$ such that $\ln p(y|\theta)$ is differentiable with*

Score Function

respect to θ. The score $V_\theta(Y)$ *of* $\mathrm{P}(y|\theta)$ *is a random variable defined by*[7]

$$V_\theta(Y) := \partial_\theta \ln p(Y|\theta) = \partial_\theta \sum_{i=1}^{m} \ln p(y_i|\theta) = \sum_{i=1}^{m} \frac{\partial_\theta p(y_i|\theta)}{p(y_i|\theta)}. \tag{3.27}$$

This score tells us how much the likelihood of the data depends on the different components of θ, and thus, in the maximum likelihood procedure, how much the data affect the choice of θ. The covariance of $V_\theta(Y)$ is called the Fisher information matrix *I. It is given by*

Fisher Information

$$I_{ij} := \mathbf{E}_\theta \left[\partial_{\theta_i} \ln p(Y|\theta) \cdot \partial_{\theta_j} \ln p(Y|\theta) \right]. \tag{3.28}$$

Covariance

and the covariance matrix *B of the estimator $\hat{\theta}(Y)$ is defined by*

$$B_{ij} := \mathbf{E}_\theta \left[\left(\hat{\theta}_i - \mathbf{E}_\theta \left[\hat{\theta}_i \right] \right) \left(\hat{\theta}_j - \mathbf{E}_\theta \left[\hat{\theta}_j \right] \right) \right]. \tag{3.29}$$

The covariance matrix B tells us the amount of variation of the estimator. It can therefore be used (e.g., by Chebychev's inequality) to bound the probability that $\hat{\theta}(Y)$ deviates from θ by more than a certain amount.

Remark 3.10 (Expected Value of Fisher Score) *One can check that the expected value of $V_\theta(Y)$ is* 0 *since*

$$\mathbf{E}_\theta [V_\theta(Y)] = \int p(Y|\theta) \partial_\theta \ln p(Y|\theta) dY = \partial_\theta \int p(Y|\theta) dY = \partial_\theta 1 = 0. \tag{3.30}$$

Average Fisher Score Vanishes

In other words, the contribution of Y to the adjustment of θ averages to 0 *over all possible Y, drawn according to* $\mathrm{P}(Y|\theta)$*. Equivalently we could say that the average likelihood for Y drawn according to* $\mathrm{P}(Y|\theta)$ *is extremal, provided we choose θ: the derivative of the expected likelihood of the data* $\mathbf{E}_\theta \left[\ln \mathrm{P}(Y|\theta) \right]$ *with respect to θ vanishes. This is also what we expect, namely that the "proper" distribution is on average the one with the highest likelihood.*

The following theorem gives a lower bound on the variance of an estimator, i.e. B is found in terms of the Fisher information I. This is useful to determine how well a given estimator performs with respect to the one with the lowest possible variance.

Theorem 3.11 (Cramér and Rao [425]) *Any unbiased estimator $\hat{\theta}(Y)$ satisfies*

$$\det IB \geq 1. \tag{3.31}$$

Proof We prove (3.31) for the scalar case. The extension to matrices is left as an exercise (see Problem 3.10). Using the Cauchy-Schwarz inequality, we obtain

$$\left(\mathbf{E}_\theta \left[(V_\theta(Y) - \mathbf{E}_\theta [V_\theta(Y)]) \left(\hat{\theta}(Y) - \mathbf{E}_\theta \left[\hat{\theta}(Y) \right] \right) \right] \right)^2 \tag{3.32}$$

$$\leq \mathbf{E}_\theta \left[(V_\theta(Y) - \mathbf{E}_\theta [V_\theta(Y)])^2 \right] \mathbf{E}_\theta \left[\left(\hat{\theta}(Y) - \mathbf{E}_\theta \left[\hat{\theta}(Y) \right] \right)^2 \right] = IB. \tag{3.33}$$

7. Recall that $\partial_\theta p(Y|\theta)$ is the gradient of $p(Y|\theta)$ with respect to the parameters $\theta_1, \ldots, \theta_n$.

At the same time, $\mathbf{E}_\theta[V_\theta(Y)] = 0$ implies that

$$\left(\mathbf{E}_\theta\left[(V_\theta(Y) - \mathbf{E}_\theta[V_\theta(Y)])\left(\hat{\theta}(Y) - \mathbf{E}_\theta\left[\hat{\theta}(Y)\right]\right)\right]\right)^2 \tag{3.34}$$

$$= \mathbf{E}_\theta\left[V_\theta(Y)\hat{\theta}(Y)\right]^2 \tag{3.35}$$

$$= \left(\int p(Y|\theta)V_\theta(Y)\hat{\theta}(Y)dY\right)^2$$

$$= \left(\partial_\theta \int p(Y|\theta)\hat{\theta}(Y)dY\right)^2 = (\partial_\theta\theta)^2 = 1, \tag{3.36}$$

since we may interchange integration by Y and ∂_θ. ■

Eq. (3.31) lends itself to the definition of a one-number summary of the properties of an estimator, namely how closely the inequality is met.

Definition 3.12 (Efficiency) *The statistical efficiency e of an estimator $\hat{\theta}(Y)$ is defined as*

$$e := 1/\det IB. \tag{3.37}$$

The closer e is to 1, the lower the variance of the corresponding estimator $\hat{\theta}(Y)$. For a special class of estimators minimizing loss functions, the following theorem allows us to compute B and e efficiently.

Theorem 3.13 (Murata, Yoshizawa, Amari [379, Lemma 3]) *Assume that $\hat{\theta}$ is defined by $\hat{\theta}(Y) := \operatorname{argmin}_\theta d(Y, \theta)$ and that d is a twice differentiable function in θ. Then asymptotically, for increasing sample size $m \to \infty$, the variance B is given by $B = Q^{-1}GQ^{-1}$. Here*

Asymptotic Variance

$$G_{ij} := \operatorname{cov}_\theta\left[\partial_{\theta_i} d(Y, \theta), \partial_{\theta_j} d(Y, \theta)\right] \text{ and} \tag{3.38}$$

$$Q_{ij} := \mathbf{E}_\theta\left[\partial^2_{\theta_i\theta_j} d(Y, \theta)\right], \tag{3.39}$$

and therefore $e = (\det Q)^2/(\det IG)$.

This means that for the class of estimators defined via d, the evaluation of their asymptotic efficiency can be conveniently achieved via (3.38) and (3.39). For scalar valued estimators $\theta(Y) \in \mathbb{R}$, these expressions can be greatly simplified to

$$I = \int \left(\partial_\theta \ln p(Y|\theta)\right)^2 d\mathrm{P}(Y|\theta), \tag{3.40}$$

$$G = \int (\partial_\theta d(Y, \theta))^2 \, d\mathrm{P}(Y|\theta), \tag{3.41}$$

$$Q = \int \partial^2_\theta d(Y, \theta) d\mathrm{P}(Y|\theta). \tag{3.42}$$

Finally, in the case of continuous densities, Theorem 3.13 may be extended to piecewise twice differentiable continuous functions d, by convolving the latter with a twice differentiable smoothing kernel, and letting the width of the smoothing kernel converge to zero. We will make use of this observation in the next section when studying the efficiency of some estimators.

The current section concludes with the proof that the maximum likelihood estimator meets the Cramér-Rao bound.

Theorem 3.14 (Efficiency of Maximum Likelihood [118, 218, 43]) *The maximum likelihood estimator (cf. (3.18) and (3.19)) given by*

$$\hat{\theta}(Y) := \operatorname*{argmax}_{\theta} \ln p(Y|\theta) = \operatorname*{argmin}_{\theta} \mathcal{L}[\theta] \tag{3.43}$$

is asymptotically efficient ($e = 1$).

To keep things simple we will prove (3.43) only for the class of twice differentiable continuous densities by applying Theorem 3.13. For a more general proof see [118, 218, 43].

Proof By construction, G is equal to the Fisher information matrix, if we choose d according to (3.43). Hence a sufficient condition is that $Q = -I$, which is what we show below. To this end we expand the integrand of (3.42),

$$\partial_\theta^2 d(Y,\theta) = \partial_\theta^2 \ln p(Y|\theta) = \frac{\partial_\theta^2 p(Y|\theta)}{p(Y|\theta)} - \left(\frac{\partial_\theta p(Y|\theta)}{p(Y|\theta)}\right)^2 = \frac{\partial_\theta^2 p(Y|\theta)}{p(Y|\theta)} - V_\theta^2(Y). \tag{3.44}$$

The expectation of the second term in (3.44) equals $-I$. We now show that the expectation of the first term vanishes;

$$\int p(Y|\theta)\frac{\partial_\theta^2 p(Y|\theta)}{p(Y|\theta)}dY = \partial_\theta^2 \int p(Y|\theta)dY = \partial_\theta^2 1 = 0. \tag{3.45}$$

Hence $Q = -I$ and thus $e = Q^2/(IG) = 1$. This proves that the maximum likelihood estimator is asymptotically efficient. ∎

It appears as if the best thing we could do is to use the maximum likelihood (ML) estimator. Unfortunately, reality is not quite as simple as that. First, the above statement holds only asymptotically. This leads to the (justified) suspicion that for finite sample sizes we may be able to do better than ML estimation. Second, practical considerations such as the additional goal of sparse decomposition may lead to the choice of a non-optimal loss function.

Finally, we may not know the true density model, which is required for the definition of the maximum likelihood estimator. We can try to make an educated guess; bad guesses of the class of densities, however, can lead to large errors in the estimation (see, e.g., [251]). This prompted the development of robust estimators.

3.4 Robust Estimators

So far, in order to make any practical predictions, we had to *assume* a certain class of distributions from which P(Y) was chosen. Likewise, in the case of risk functionals, we also assumed that training and test data are identically distributed. This section provides tools to safeguard ourselves against cases where the above

assumptions are not satisfied.

Outliers

More specifically, we would like to avoid a certain fraction ν of 'bad' observations (often also referred to as 'outliers') seriously affecting the quality of the estimate. This implies that the influence of individual patterns should be bounded from above. Huber [250] gives a detailed list of desirable properties of a robust estimator. We refrain from reproducing this list at present, or committing to a particular definition of robustness.

As usual for the estimation of location parameter context (i.e. estimation of the expected value of a random variable) we assume a specific parametric form of $p(Y|\theta)$, namely

$$p(Y|\theta) = \prod_{i=1}^{m} p(y_i|\theta) = \prod_{i=1}^{m} p(y_i - \theta). \tag{3.46}$$

Unless stated otherwise, this is the formulation we will use throughout this section.

3.4.1 Robustness via Loss Functions

Huber's idea [250] in constructing a robust estimator was to take a loss function as provided by the maximum likelihood framework, and modify it in such a way as to limit the influence of each individual pattern. This is done by providing an upper bound on the slope of $-\ln p(Y|\theta)$. We shall see that methods such as the trimmed mean or the median are special cases thereof. The ε-insensitive loss function can also be viewed as a trimmed estimator. This will lead to the development of adaptive loss functions in the subsequent sections. We begin with the main theorem of this section.

Mixture Densities

Theorem 3.15 (Robust Loss Functions (Huber [250])) *Let $\mathfrak{P}$ be a class of densities formed by*

$$\mathfrak{P} := \{p|p = (1-\varepsilon)p_0 + \varepsilon p_1\} \text{ where } \varepsilon \in (0,1) \text{ and } p_0 \text{ are known.} \tag{3.47}$$

Moreover assume that both p_0 and p_1 are symmetric with respect to the origin, their logarithms are twice continuously differentiable, $\ln p_0$ is convex and known, and p_1 is unknown. Then the density

$$\bar{p}(\theta) := (1-\varepsilon) \begin{cases} p_0(\theta) & \text{if } |\theta| \leq \theta_0 \\ p_0(\theta_0)e^{-k(|\theta|-\theta_0)} & \text{otherwise} \end{cases} \tag{3.48}$$

is robust in the sense that the maximum likelihood estimator corresponding to (3.48) has minimum variance with respect to the "worst" possible density $p_{\text{worst}} = (1-\varepsilon)p_0 + \varepsilon p_1$: it is a saddle point (located at p_{worst}) in terms of variance with respect to the true density $p \in \mathfrak{P}$ and the density $\bar{p} \in \mathfrak{P}$ used in estimating the location parameter. This means that no density p has larger variance than p_{worst} and that for $p = p_{\text{worst}}$ no estimator is better than the one where $\bar{p} = p_{\text{worst}}$, as used in the robust estimator.

The constants $k > 0$ and θ_0 are obtained by the normalization condition, that $\bar{p}$ be a

proper density and that the first derivative in $\ln \bar{p}$ *be continuous.*

Proof To show that $\bar{p}$ is a saddle point in $\mathfrak{P}$ we have to prove that (a) no estimation procedure other than the one using $\ln \bar{p}$ as the loss function has lower variance for the density $\bar{p}$, and that (b) no density has higher variance than $\bar{p}$ if $\ln \bar{p}$ is used as loss function. Part (a) follows immediately from the Cramér-Rao theorem (Th. 3.11); part (b) can be proved as follows.

We use Theorem 3.13, and a proof technique pointed out in [559], to compute the variance of an estimator using $\ln \bar{p}$ as loss function;

$$B = \frac{\int \left(\partial_\theta \ln \bar{p}(y|\theta)\right)^2 \left((1-\varepsilon)p_0(y|\theta) + \varepsilon p'(y|\theta)\right) dy}{\int \partial_\theta^2 \ln \bar{p}(y|\theta) \left((1-\varepsilon)p_0(y|\theta) + \varepsilon p'(y|\theta)\right) dy}. \tag{3.49}$$

Here p' is an arbitrary density which we will choose such that B is maximized. By construction,

$$\left(\partial_\theta \ln \bar{p}(y|\theta)\right)^2 = \begin{cases} \left(\partial_\theta \ln p_0(y|\theta)\right)^2 \leq k^2 & \text{if } |y-\theta| \leq \theta_0, \\ k^2 & \text{otherwise,} \end{cases} \tag{3.50}$$

$$\partial_\theta^2 \ln \bar{p}(y|\theta) = \begin{cases} \partial_\theta^2 \ln p_0(y|\theta) \geq 0 & \text{if } |y-\theta| \leq \theta_0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.51}$$

Thus any density p' which is 0 in $[-\theta_0, \theta_0]$ will minimize the denominator (the term depending on p' will be 0, which is the lowest obtainable value due to (3.51)), and maximize the numerator, since in the latter the contribution of p' is always limited to $k^2\varepsilon$. Now $\varepsilon^{-1}\left(\bar{p} - (1-\varepsilon)p_0\right)$ is exactly such a density. Hence the saddle point property holds. ■

Remark 3.16 (Robustness Classes) *If we have more knowledge about the class of densities* $\mathfrak{P}$*, a different loss function will have the saddle point property. For instance, using a similar argument as above, one can show that the normal distribution is robust in the class of all distributions with bounded variance. This implies that among all possible distributions with bounded variance, the estimator of the mean of a normal distribution has the highest variance.*

Likewise, the Laplacian distribution is robust in the class of all symmetric distributions with density $p(0) \geq c$ *for some fixed* $c > 0$ *(see [559, 251] for more details).*

Hence, even though a loss function defined according to Theorem 3.15 is generally desirable, we may be less cautious, and use a different loss function for improved performance, when we have additional knowledge of the distribution.

Remark 3.17 (Mean and Median) *Assume we are dealing with a mixture of a normal distribution with variance* σ^2 *and an additional unknown distribution with weight at most* ε*. It is easy to check that the application of Theorem 3.15 to normal distributions yields Huber's robust loss function from Table 3.1.*

The maximizer of the likelihood (see also Problem 3.17) is a trimmed mean estimator which discards ε *of the data: effectively all* θ_i *deviating from the mean by more than* σ *are*

ignored and the mean is computed from the remaining data. Hence Theorem 3.15 gives a formal justification for this popular type of estimator.

If we let $\varepsilon \to 1$ we recover the median estimator which stems from a Laplacian distribution. Here, all patterns but the median one are discarded.

Trimmed Interval Estimator

Besides the classical examples of loss functions and density models, we might also consider a slightly unconventional estimation procedure: use the average between the k-smallest and the k-largest of all observations θ observations as the estimated mean of the underlying distribution (for sorted observations θ_i with $\theta_i \leq \theta_j$ for $1 \leq i \leq j \leq m$ the estimator computes $(\theta_k + \theta_{m-k+1})/2$). This procedure makes sense, for instance, when we are trying to infer the mean of a random variable generated by roundoff noise (i.e., noise whose density is constant within some bounded interval) plus an additional unknown amount of noise.

Support Patterns

Note that both the patterns strictly *inside* or *outside* an interval of size $[-\varepsilon, \varepsilon]$ around the estimate have no direct influence on the outcome. Only patterns *on* the boundary matter. This is a very similar situation to the behavior of Support Vector Machines in regression, and one can show that it corresponds to the minimizer of the ε-insensitive loss function (3.9). We will study the properties of the latter in more detail in the following section and thereafter show how it can be transformed into an adaptive risk functional.

3.4.2 Efficiency and the ε-Insensitive Loss Function

The tools of Section 3.3.2 allow us to analyze the ε-insensitive loss function in more detail. Even though the asymptotic estimation of a location parameter setting is a gross oversimplification of what is happening in a SV regression estimator (where we estimate a nonparametric function, and moreover have only a limited number of observations at our disposition), it will provide us with useful insights into this more complex case [510, 481].

In a first step, we compute the efficiency of an estimator, for several noise models and amounts of variance, using a density corresponding to the ε-insensitive loss function (cf. Table 3.1);

$$p_\varepsilon(y|\theta) = \frac{1}{2+2\varepsilon} \exp(-|y-\theta|_\varepsilon) = \frac{1}{2+2\varepsilon} \begin{cases} 1 & \text{if } |y-\theta| \leq \varepsilon, \\ \exp(\varepsilon - |y-\theta|) & \text{otherwise.} \end{cases} \tag{3.52}$$

For this purpose we have to evaluate the quantities G (3.41) and Q (3.42) of Theorem 3.13. We obtain

$$G = m \int \left(\partial_\theta \ln p(y|\theta)\right)^2 d\mathrm{P}(y|\theta) = m \left(1 - \int_{-\varepsilon}^{\varepsilon} p(y|\theta) dy\right), \tag{3.53}$$

$$Q = m \int \partial_\theta^2 \ln p(y|\theta) d\mathrm{P}(y|\theta) = m \left(p(-\varepsilon+\theta|\theta) + p(\varepsilon+\theta|\theta)\right). \tag{3.54}$$

The Fisher information I of m iid random variables distributed according to p_θ is m-times the value of a single random variable. Thus all dependencies on m in e cancel out and we can limit ourselves to the case of $m = 1$ for the analysis of the

efficiency of estimators.

Now we may check what happens if we use the ε-insensitive loss function for different types of noise model. For the sake of simplicity we begin with Gaussian noise.

Example 3.18 (Gaussian Noise) *Assume that y is normally distributed with zero mean (i.e. $\theta = 0$) and variance σ. By construction, the minimum obtainable variance is $I^{-1} = \sigma^2$ (recall that $m = 1$). Moreover (3.53) and (3.54) yield*

$$\frac{G}{Q^2} = \sigma^2 \exp\left(\frac{\varepsilon^2}{\sigma^2}\right)\left(1 - \operatorname{erf}\frac{\varepsilon}{\sqrt{2}\sigma}\right). \tag{3.55}$$

The efficiency $e = \frac{Q^2}{GI}$ is maximized for $\varepsilon = 0.6120\sigma$. This means that if the underlying noise model is Gaussian with variance σ and *we have to use an ε-insensitive loss function to estimate a location parameter, the most efficient estimator from this family is given by $\varepsilon = 0.6120\sigma$.*

The consequence of (3.55) is that the optimal value of ε scales linearly with σ. Of course, we could just use squared loss in such a situation, but in general, we will not know the exact noise model, and squared loss does not lead to robust estimators. The following lemma (which will come handy in the next section) shows that this is a general property of the ε-insensitive loss.

Lemma 3.19 (Linear Dependency between ε-Tube Width and Variance) *Denote by p a symmetric density with variance $\sigma > 0$. Then the optimal value of ε (i.e. the value that achieves maximum asymptotic efficiency) for an estimator using the ε-insensitive loss is given by*

$$\varepsilon_{\text{opt}} = \sigma \operatorname*{argmin}_{\tau} \frac{1}{\left(p_{\text{std}}(-\tau) + p_{\text{std}}(\tau)\right)^2}\left(1 - \int_{-\tau}^{\tau} p_{\text{std}}(\tau')d\tau'\right), \tag{3.56}$$

where $p_{\text{std}}(\tau) := \sigma p(\sigma\tau + \theta|\theta)$ is the standardized version of $p(y|\theta)$, i.e. it is obtained by rescaling $p(y|\theta)$ to zero mean and unit variance.

Since p_{std} is independent of σ, we have a linear dependency between ε_{opt} and σ. The scaling factor depends on the noise model.

Proof We prove (3.56) by rewriting the efficiency $e(\varepsilon)$ in terms of p_{std} via $p(y|\theta) = \sigma^{-1} p_{\text{std}}(\sigma^{-1}(y - \theta))$. This yields

$$e(\varepsilon) = \frac{Q^2}{IG} = \frac{\left(\sigma^{-1} p_{\text{std}}(-\sigma^{-1}\varepsilon) + \sigma^{-1} p_{\text{std}}(\sigma^{-1}\varepsilon)\right)^2}{\sigma^{-2}\left(1 - \int_{-\varepsilon}^{\varepsilon} \sigma^{-1} p_{\text{std}}(\sigma^{-1}\theta)d\theta\right)} = \frac{\left(p_{\text{std}}(-\sigma^{-1}\varepsilon) + p_{\text{std}}(\sigma^{-1}\varepsilon)\right)^2}{\left(1 - \int_{-\sigma^{-1}\varepsilon}^{\sigma^{-1}\varepsilon} p_{\text{std}}(\theta)d\theta\right)}$$

The maximum of $e(\varepsilon)$ does not depend directly on ε, but on $\sigma^{-1}\varepsilon$ (which is independent of σ). Hence we can find $\operatorname{argmax}_{\varepsilon} e(\varepsilon)$ by solving (3.56). ■

Lemma 3.19 made it apparent that in order to adjust ε we have to know σ beforehand. Unfortunately, the latter is usually unknown at the beginning of the

estimation procedure.[8] The solution to this dilemma is to make ε adaptive.

3.4.3 Adaptive Loss Functions

We again consider the trimmed mean estimator, which discards a predefined fraction of largest and smallest samples. This method belongs to the more general class of quantile estimators, which base their estimates on the value of samples in a certain quantile. The latter methods do not require prior knowledge of the variance, and adapt to whatever scale is required. What we need is a technique which connects σ (in Huber's robust loss function) or ε (in the ε-insensitive loss case) with the deviations between the estimate $\hat{\theta}$ and the random variables y_i.

Let us analyze what happens to the negative log likelihood, if, in the ε-insensitive case, we change ε to $\varepsilon + \delta$ (with $\delta \in \mathbb{R}$) while keeping $\hat{\theta}$ fixed. In particular we assume that $|\delta|$ is chosen sufficiently small such that for all $i = 1, \ldots, m$,

$$|\hat{\theta} - y_i| \begin{cases} \leq \varepsilon + \delta & \text{if } |\hat{\theta} - y_i| < \varepsilon \\ \geq \varepsilon + \delta & \text{if } |\hat{\theta} - y_i| > \varepsilon \end{cases} \tag{3.57}$$

Moreover denote by $m_<, m_=, m_>$ the number of samples for which $|\hat{\theta} - y_i|$ is less than, equal to, or greater than ε, respectively. Then

$$\begin{aligned} \sum_{i=1}^{m} |\hat{\theta} - y_i|_{\varepsilon+\delta} &= \sum_{|\hat{\theta}-y_i|<\varepsilon} |\hat{\theta} - y_i|_\varepsilon + \sum_{|\hat{\theta}-y_i|>\varepsilon} |\hat{\theta} - y_i|_\varepsilon - m_>\delta + \sum_{|\hat{\theta}-y_i|=\varepsilon} |\hat{\theta} - y_i|_{\varepsilon+\delta} \\ &= \sum_{i=1}^{m} |\hat{\theta} - y_i|_\varepsilon - \begin{cases} m_>\delta & \text{if } \delta > 0, \\ (m_< + m_=)\delta & \text{otherwise.} \end{cases} \end{aligned} \tag{3.58}$$

In other words, the amount by which the loss changes depends only on the quantiles at ε. What happens if we make ε itself a variable of the optimization problem? By the scaling properties of (3.58) one can see that for $\nu \in [0, 1]$

$$\underset{\hat{\theta},\varepsilon}{\text{minimize}} \frac{1}{m} \sum_{i=1}^{m} |\hat{\theta} - y_i|_\varepsilon - \nu\varepsilon \tag{3.59}$$

ν-Property

is minimized if ε is chosen such that

$$\frac{m_>}{m} \leq \nu \leq \frac{m_> + m_=}{m}. \tag{3.60}$$

This relation holds since at the solution $(\hat{\theta}, \varepsilon)$ the solution also has to be optimal wrt. ε alone while keeping $\hat{\theta}$ fixed. In the latter case, however, the derivatives of

8. The obvious question is why one would ever like to choose an ε-insensitive loss in the presence of Gaussian noise in the first place. If the complexity of the function expansion is of no concern and the highest accuracy is required, squared loss is to be preferred. In most cases, however, it is not quite clear what *exactly* the type of the additive noise model is. This is when we would like to have a more conservative estimator. In practice, the ε-insensitive loss has been shown to work rather well on a variety of tasks (Chapter 9).

the log-likelihood (i.e. error) term wrt. ε at the solution are given by $\frac{m_>}{m}$ and $\frac{m_>+m_=}{m}$ on the left and right hand side respectively.[9] These have to cancel with ν which proves the claim. Furthermore, computing the derivative of (3.59) with respect to $\hat{\theta}$ shows that the number of samples outside the interval $[\theta-\varepsilon, \theta+\varepsilon]$ has to be equal on both halves $(-\infty, \theta-\varepsilon)$ and $(\theta+\varepsilon, \infty)$. We have the following theorem:

Theorem 3.20 (Quantile Estimation as Optimization Problem [481]) *A quantile procedure to estimate the mean of a distribution by taking the average of the samples at the $\frac{\nu}{2}$th and $(1-\frac{\nu}{2})$th quantile is equivalent to minimizing (3.59). In particular,*

1. *ν is an upper bound on the fraction of samples outside the interval $[\theta-\varepsilon, \theta+\varepsilon]$.*
2. *ν is a lower bound on the fraction of samples outside the interval $]\theta-\varepsilon, \theta+\varepsilon[$.*
3. *If the distribution $p(\theta)$ is continuous, for all $\nu \in [0,1]$*

$$\lim_{m\to\infty} \mathrm{P}\left\{\frac{m_=}{m} < \varepsilon\right\} = 1 \text{ for all } \varepsilon > 0. \tag{3.61}$$

One might question the practical advantage of this method over direct trimming of the sample Y. In fact, the use of (3.59) is not recommended if all we want is to estimate θ. That said, (3.59) does allow us to employ trimmed estimation in the nonparametric case, cf. Chapter 9.

Extension to General Robust Estimators

Unfortunately, we were unable to find a similar method for Huber's robust loss function, since in this case the change in the negative log-likelihood incurred by changing σ not only involves the (statistical) rank of y_i, but also the exact location of samples with $|y_i - \theta| < \sigma$.

One way to overcome this problem is re-estimate σ adaptively while minimizing a term similar to (3.59) (see [180] for details in the context of boosting, Section 10.6.3 for a discussion of online estimation techniques, or [251] for a general overview).

3.4.4 Optimal Choice of ν

Let us return to the ε-insensitive loss. A combination of Theorems 3.20, 3.13 and Lemma 3.19 allows us to compute optimal values of ν for various distributions, provided that an ε-insensitive loss function is to be used in the estimation procedure.[10]

The idea is to determine the optimal value of ε for a fixed density $p(y|\theta)$ via (3.56), and compute the corresponding fraction ν of patterns outside the interval $[-\varepsilon+\theta, \varepsilon+\theta]$.

9. Strictly speaking, the derivative is not defined at ε; the lhs and rhs values are defined, however, which is sufficient for our purpose.

10. This is not optimal in the sense of Theorem 3.15, which suggests the use of a more adapted loss function. However (as already stated in the introduction of this chapter), algorithmic or technical reasons such as computationally efficient solutions or limited memory may provide sufficient motivation to use such a loss function.

Table 3.2 Optimal ν and ε for various degrees of polynomial additive noise.

Polynomial Degree d	1	2	3	4	5
Optimal ν	1	0.5405	0.2909	0.1898	0.1384
Optimal ε for unit variance	0	0.6120	1.1180	1.3583	1.4844
Polynomial Degree d	6	7	8	9	10
Optimal ν	0.1080	0.0881	0.0743	0.0641	0.0563
Optimal ε for unit variance	1.5576	1.6035	1.6339	1.6551	1.6704

Theorem 3.21 (Optimal Choice of ν) *Denote by p a symmetric density with variance $\sigma > 0$ and by p_{std} the corresponding rescaled density with zero mean and unit variance. Then the optimal value of ν (i.e. the value that achieves maximum asymptotic efficiency) for an estimator using the ε-insensitive loss is given by*

$$\nu = 1 - \int_{-\varepsilon}^{\varepsilon} p_{\text{std}}(y)dy \tag{3.62}$$

where ε is chosen according to (3.56). This expression is independent of σ.

Proof The independence of σ follows from the fact that ν depends only on p_{std}. Next we show (3.62). For a given density p, the asymptotically optimal value of ε is given by Lemma 3.19. The average fraction of patterns outside the interval $[\hat{\theta} - \varepsilon_{\text{opt}}, \hat{\theta} + \varepsilon_{\text{opt}}]$ is

$$\nu = 1 - \int_{-\varepsilon_{\text{opt}}+\theta}^{\varepsilon_{\text{opt}}+\theta} p(y|\theta)dy = 1 - \int_{-\sigma^{-1}\varepsilon_{\text{opt}}}^{\sigma^{-1}\varepsilon_{\text{opt}}} p_{\text{std}}(y)dy, \tag{3.63}$$

which depends only on $\sigma^{-1}\varepsilon_{\text{opt}}$ and is thus independent of σ. Combining (3.63) with (3.56) yields the theorem. ∎

This means that given the *type* of additive noise, we *can* determine the value of ν such that it yields the asymptotically most efficient estimator *independent* of the *level* of the noise. These theoretical predictions have since been confirmed rather accurately in a set of regression experiments [95].

Let us now look at some special cases.

Example 3.22 (Optimal ν for Polynomial Noise) *Arbitrary polynomial noise models ($\propto e^{-|\theta|^d}$) with unit variance can be written as*

$$p(y) = c_p \exp\left(-c'_p|y|^p\right) \text{ where } c_p = \frac{1}{2}\sqrt{\frac{\Gamma(3/d)}{\Gamma(1/d)}}\frac{d}{\Gamma(1/d)} \text{ and } c'_p = \left(\sqrt{\frac{\Gamma(3/d)}{\Gamma(1/d)}}\right)^d,$$

where $\Gamma(x)$ is the gamma function. Figure 3.3 shows ν_{opt} for polynomial degrees in the interval $[1, 10]$. For convenience, the explicit numerical values are repeated in Table 3.2.

Heavy Tails → Large ν

Observe that as the distribution becomes "lighter-tailed", the optimal ν decreases; in other words, we may then use a larger amount of the data for the purpose of estimation. This is reasonable since it is only for very long tails of the distribution (data with many

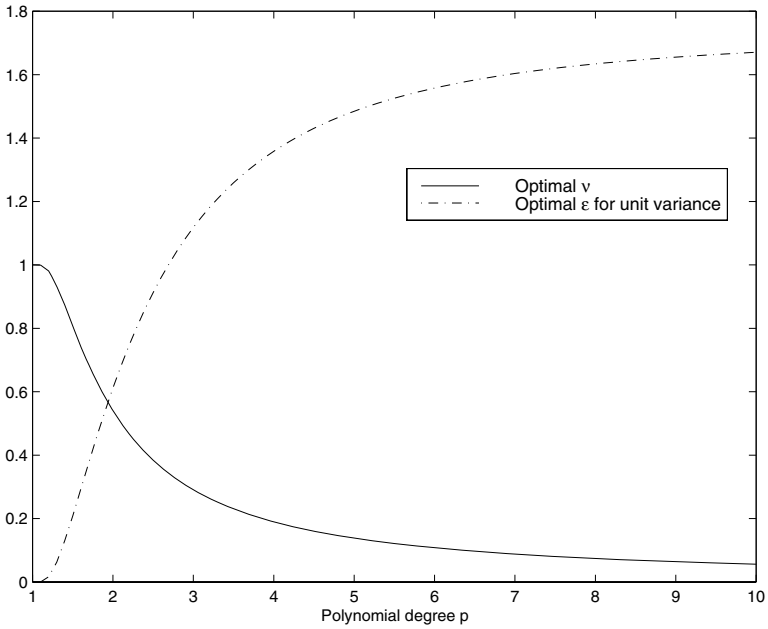


Figure 3.3 Optimal ν and ε for various degrees of polynomial additive noise.

outliers) that we have to be conservative and discard a large fraction of observations.

Even though we derived these relations solely for the case where a single number (θ) has to be estimated, experiments show that the same scaling properties hold for the nonparametric case. It is still an open research problem to establish this connection exactly.

As we shall see, in the nonparametric case, the effect of ν will be that it both determines the number of Support Vectors (i.e., the number of basis functions needed to expand the solution) and also the fraction of function values $f(x_i)$ with deviation larger than ε from the corresponding observations. Further information on this topic, both from the statistical and the algorithmic point of view, can be found in Section 9.3.

3.5 Summary

We saw in this chapter that there exist two complementary concepts as to how risk and loss functions should be designed. The first one is data driven and uses the incurred loss as its principal guideline, possibly modified in order to suit the need of numerical efficiency. This leads to loss functions and the definitions of empirical and expected risk.

A second method is based on the idea of estimating (or at least approximating) the distribution which may be responsible for generating the data. We showed

that in a Maximum Likelihood setting this concept is rather similar to the notions of risk and loss, with $c(x, y, f(x)) = -\ln p(y|x, f(x))$ as the link between both quantities.

This point of view allowed us to analyze the properties of estimators in more detail and provide lower bounds on the performance of unbiased estimators, i.e. the Cramér-Rao theorem. The latter was then used as a benchmarking tool for various loss functions and density models, such as the ε-insensitive loss. The consequence of this analysis is a corroboration of experimental findings that there exists a linear correlation between the amount of noise in the observations and the optimal width of ε.

This, in turn, allowed us to construct adaptive loss functions which adjust themselves to the amount of noise, much like trimmed mean estimators. These formulations can be used directly in mathematical programs, leading to ν-SV algorithms in subsequent chapters. The question of which choices are optimal in a finite sample size setting remains an open research problem.

3.6 Problems

3.1 (Soft Margin and Logistic Regression •) *The soft margin loss function c_{soft} and the logistic loss c_{logist} are asymptotically almost the same; show that*

$$\lim_{f\to\infty} \left(c_{\text{soft}}(x, 1, f) - c_{\text{logist}}(x, 1, f)\right) = 1 \tag{3.64}$$

$$\lim_{f\to-\infty} \left(c_{\text{soft}}(x, 1, f) - c_{\text{logist}}(x, 1, f)\right) = 0. \tag{3.65}$$

3.2 (Multi-class Discrimination ••) *Assume you have to solve a classification problem with M different classes. Discuss how the number of functions used to solve this task affects the quality of the solution.*

- *How would the loss function look if you were to use only one real-valued function $f : \mathcal{X} \to \mathbb{R}$. Which symmetries are violated in this case (hint: what happens if you permute the classes)?*
- *How many functions do you need if each of them makes a binary decision $f : \mathcal{X} \to \{0, 1\}$?*
- *How many functions do you need in order to make the solution permutation symmetric with respect to the class labels?*
- *How should you assess the classification error? Is it a good idea to use the misclassification rate of one individual function as a performance criterion (hint: correlation of errors)? By how much can this error differ from the total misclassification error?*

3.3 (Mean and Median •) *Assume 8 people want to gather for a meeting; 5 of them live in Stuttgart and 3 in Munich. Where should they meet if (a) they want the total distance traveled by all people to be minimal, (b) they want the average distance traveled per person to be minimal, or (c) they want the average squared distance to be minimal? What happens*

to the meeting points if one of the 3 *people moves from Munich to Sydney?*

3.4 (Locally Adaptive Loss Functions ●●●) *Assume that the loss function* $c(x, y, f(x))$ *varies with* x*. What does this mean for the expected loss? Can you give a bound on the latter even if you know* $p(y|x)$ *and* f *at every point but know* c *only on a finite sample (hint: construct a counterexample)? How will things change if* c *cannot vary much with* x*?*

3.5 (Transduction Error ●●●) *Assume that we want to minimize the test error of misclassification* $R_{\text{test}}[f]$*, given a training sample* $\{(x_1, y_1), \ldots, (x_m, y_m)\}$*, a test sample* $\{x'_1, \ldots, x'_{m'}\}$ *and a loss function* $c(x, y, f(x))$*.*

Show that any loss function $c'(x', f(x'))$ *on the test sample has to be symmetric in* f*, i.e.* $c'(x', f(x')) = c'(x', -f(x'))$*. Prove that no non-constant convex function can satisfy this property. What does this mean for the practical solution of optimization problem? See [267, 37, 211, 103] for details.*

3.6 (Convexity and Uniqueness ●●) *Show that the problem of estimating a location parameter (a single scalar) has an interval* $[a, b] \subset \mathbb{R}$ *of equivalent global minima if the loss functions are convex. For non-convex loss functions construct an example where this is not the case.*

3.7 (Linearly Dependent Parameters ●●) *Show that in a linear model* $f = \sum_i \alpha_i f_i$ *on* $\mathcal{X}$ *it is impossible to find a unique set of optimal parameters* α_i *if the functions* f_i *are not linearly independent. Does this have any effect on* f *itself?*

3.8 (Ill-posed Problems ●●●) *Assume you want to solve the problem* $Ax = y$ *where* A *is a symmetric positive definite matrix, i.e., a matrix with nonnegative eigenvalues. If you change* y *to* y'*, how much will the solution* x' *of* $Ax' = y'$ *differ from* x'*. Give lower and upper bounds on this quantity. Hint: decompose* y *into the eigensystem of* A*.*

3.9 (Fisher Map [258] ●●) *Show that the map*

$$U_\theta(x) := I^{-\frac{1}{2}} \partial_\theta \ln p(x|\theta) \tag{3.66}$$

maps x *into vectors with zero mean and unit variance. Chapter 13 will use this map to design kernels.*

3.10 (Cramér-Rao Inequality for Multivariate Estimators ●●) *Prove equation (3.31). Hint: start by applying the Cauchy-Schwarz inequality to*

$$\left(\det E_{\hat{\theta}}[(\hat{\theta}(\theta) - E_{\hat{\theta}}\hat{\theta}(\theta))(T_\theta(\theta) - E_{\hat{\theta}} T_\theta(\theta))^\top]\right) \tag{3.67}$$

to obtain I *and* B *and compute the expected value coefficient-wise.*

3.11 (Soft Margin Loss and Conditional Probabilities [521] ●●●) *What is the conditional probability* $p(y|x)$ *corresponding to the soft margin loss function* $c(x, y, f(x)) = \max(0, 1 - yf(x))$*?*

- *How can you fix the problem that the probabilities $p(-1|x)$ and $p(1|x)$ have to sum up to 1?*
- *How does the introduction of a third class ("don't know") change the problem? What is the problem with this approach? Hint: What is the behavior for large $|f(x)|$?*

3.12 (Label Noise ••) *Denote by $\mathrm{P}(y = 1|f(x))$ and $\mathrm{P}(y = -1|f(x))$ the conditional probabilities of labels ± 1 for a classifier output $f(x)$. How will P change if we randomly flip labels with $\eta \in (0, 1)$ probability? How should you adapt your density model?*

3.13 (Unbiased Estimators ••) *Prove that the least mean square estimator is unbiased for arbitrary symmetric distributions. Can you extend the result to arbitrary symmetric losses?*

3.14 (Efficiency of Huber's Robust Estimator ••) *Compute the efficiency of Huber's Robust Estimator in the presence of pure Gaussian noise with unit variance.*

3.15 (Influence and Robustness •••) *Prove that for robust estimators using (3.48) as their density model, the maximum change in the minimizer of the empirical risk is bounded by $\frac{\delta k}{m}$ if a sample θ_i is changed to $\theta_i + \delta$. What happens in the case of Gaussian density models (i.e., squared loss)?*

3.16 (Robustness of Gaussian Distributions [559] •••) *Prove that the normal distribution with variance σ^2 is robust among the class of distributions with bounded variance (by σ^2). Hint: show that we have a saddle point analogous to Theorem 3.15 by exploiting Theorems 3.13 and Theorem 3.14.*

3.17 (Trimmed Mean ••) *Show that under the assumption of an unknown distribution contributing at most ε, Huber's robust loss function for normal distributions leads to a trimmed mean estimator which discards ε of the data.*

3.18 (Optimal ν for Gaussian Noise •) *Give an explicit solution for the optimal ν in the case of additive Gaussian noise.*

3.19 (Optimal ν for Discrete Distribution ••) *Assume that we have a noise model with a discrete distribution of θ, where $\mathrm{P}(\theta = \epsilon) = \mathrm{P}(\theta = -\epsilon) = p_1$, $\mathrm{P}(\theta = 2\epsilon) = \mathrm{P}(\theta = -2\epsilon) = p_2$, $2(p_1 + p_2) = 1$, and $p_1, p_2 \geq 0$. Compute the optimal value of ν.*