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*

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

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

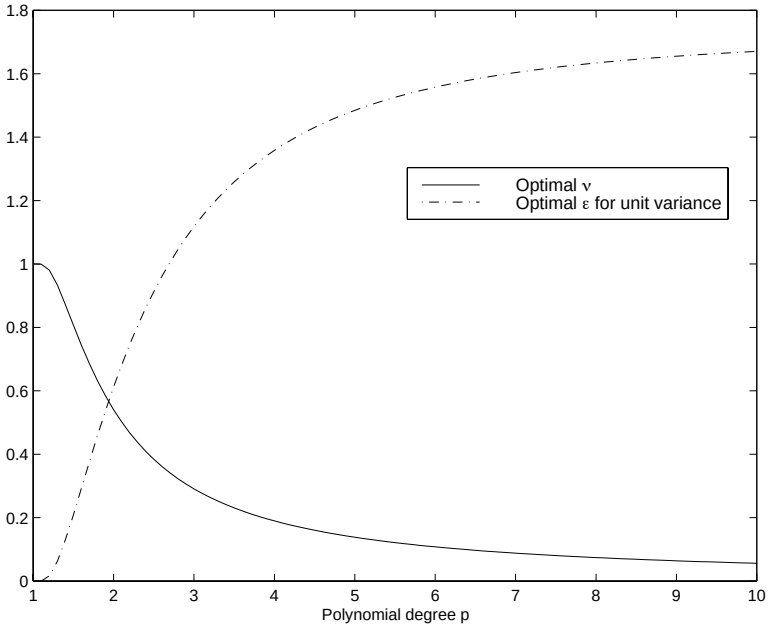

Figure 3.3 Optimal ν and ε for various degrees of polynomial additive noise.

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