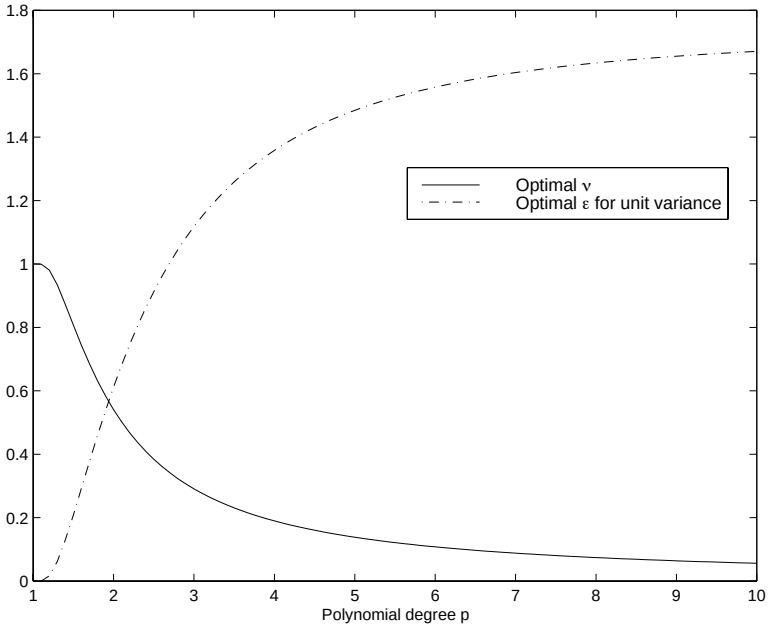

Figure 3.3 Optimal ν and ε for various degrees of polynomial additive noise.

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*

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

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

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*