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*

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*

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

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.*