

# Short Course on Optimization Methods for Statistics and Selected Topics in Classification

Lutz Dümbgen  
(University of Berne)

UC Louvain-la-Neuve, Belgium  
January 21-22, 2008

**Abstract.** In the first part I will present some tools from optimization theory which I found useful in various statistical applications. My impression is that statisticians tend to use rather ad hoc and sometimes sub-optimal methods for the optimization problems we encounter. Alternatively we sometimes use software tools for optimization as black boxes without sufficient background knowledge to understand their potential benefits and pitfalls. The aim of the first part of my lectures is to present some methods for convex optimization and to illustrate them with applications to inverse problems, mixture models and shape-constrained estimation.

In the second part I will recall the setting of classification procedures and remind you of some standard results and classifiers. It will turn out that the contents of the first part are relevant here as well. Thereafter I will try to convince you that there is an alternative approach to classification via p-values. The latter concept will be treated in some detail with an outlook for further research.

## 1 Some Tools from Optimization Theory for Statistical Inference

### 1.1 A Case Study

#### **An Inverse Problem from Physics.**

Description of the experiment and inverse problem.

An inversion formula and its quality.

#### **Maximum-Likelihood Estimation.**

Embedding the current example into mixture models.

The rationale behind ML estimation.

#### **A First Solution via the EM Algorithm.**

The general idea of the Expectation-Maximization algorithm.

Its implementation for discrete mixture models.

## **1.2 Newton and Quasi-Newton Methods**

### **Newton's Method and its Speed.**

Superlinear and quadratic convergence

Comparison with fixed-point iterations.

### **Quadratic Approximations.**

Minimizing a function via local quadratic approximations

Gradient, Newton- and quasi-Newton methods

Step size corrections.

## **1.3 Constrained Convex Optimization Problems**

### **The Setting and Characterizations of Solutions.**

Convex polytopes and cones.

Directional derivatives.

### **Log-Barrier Methods.**

A first brute force solution to constrained programming;  
good for finding starting values of more sophisticated methods.

### **Quasi-Newton Methods.**

A second class of solutions to constrained programming;  
iterative algorithms with guaranteed convergence.

### **Active-Set Methods**

A third class of solutions to constrained programming;  
convergence after finitely many steps, if target functional is quadratic.

Example: concave regression.

## **1.4 Log-Concave Density Estimation**

### **Definitions and Motivation.**

Log-concave densities as natural link between parametric and nonparametric models.

Unimodality.

### **ML Estimation.**

Existence and uniqueness of the solution.

Two characterizations of the solution.

### **Computing the MLE via an Active Set Method.**

Brief description of our implementation.

Illustration with some examples.

### **Consistency.**

Uniform rates of convergence for the log-density MLE.

Heuristical derivation of these rates.

## **2 Statistical Inference in Classification**

General setting: “Observations”  $(X, Y)$  consisting of an observed feature vector  $X \in \mathcal{X}$  and a hidden class Label  $Y \in \{1, 2, \dots, L\}$ .

Goal: Inference about  $Y$  based on  $X$  (plus training data).

### **2.1 Classifiers**

#### **Classification in an idealized setting.**

Posterior distributions.

Risk measures and Bayes-optimal classifiers.

#### **Classification with Training Data.**

Conditional and unconditional risk measures.

Various modes of cross-validation.

#### **Examples of classifiers.**

- Model-based classifiers (LDA, QDA, KNN)
- Regression-type estimators
- Kernel methods
- Regularization

## 2.2 P-Values for Classification

### From Classifiers to P-Values.

Various arguments why neither a classifier nor a posterior distribution is entirely satisfying.

### Optimal P-Values in the Idealized Setting.

Bayes-optimal p-values.

### Training Data.

Two notions of validity.

Analyzing and visualizing the training data.

Typicality indices.

Nonparametric p-values via permutation tests.

Computational issues.

Examples.

### Some Asymptotics in the Classical Framework.

Asymptotic multiple-use validity and optimality as  $n \rightarrow \infty$ .

### Outlook.

High-dimensional models and stability.

Regularization.

## References

- [1] BARTLETT, P.L., JORDAN, M.I. and MCAULIFFE, J.D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* **101**, 138–156.
- [2] A. BEN-TAL and A. NEMIROVSKI (2001). *Lectures on Modern Convex Optimization. Analysis, Algorithms, and Engineering Applications*. SIAM, Philadelphia
- [3] DÜMBGEN, L., S. FREITAG and G. JONGBLOED (2006). Estimating a unimodal distribution from interval-censored data. *J. Amer. Statist. Assoc.* **101**, 1094–1106.
- [4] DÜMBGEN, L., A. HÜSLER and K. RUFIBACH (2007). Active set and EM algorithms for log-concave densities based on complete and censored data. Technical Report 61, IMSV, University of Bern (<http://arxiv.org/abs/0707.4643>)
- [5] DÜMBGEN, L., B.-W. IGL and A. MUNK (2008). P-values for classification. Preprint (<http://arxiv.org/abs/0801.2934>).
- [6] DÜMBGEN, L. and K. RUFIBACH (2007, revised 2008). Maximum likelihood estimation of a log-concave density and its distribution function: basic properties and uniform consistency. Technical Report 66, IMSV, University of Bern (<http://arxiv.org/abs/0709.0334>)
- [7] R. FLETCHER (1987). *Practical Methods of Optimization (2nd edition)*. Wiley, New York
- [8] P.E. GILL, W. MURRAY and M.H. WRIGHT (1981). *Practical Optimization*. Academic Press
- [9] P. GROENEBOOM, G. JONGBLOED and J.A. WELLNER (2008). The support reduction algorithm for computing nonparametric function estimates in mixture models. *Scand. J. Statist.*, to appear.
- [10] MCLACHLAN, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.
- [11] ROCKAFELLAR, R.T. (1997). *Convex Analysis*. Princeton University Press, Princeton, NJ.
- [12] SCHÖLKOPF, B. and A. SMOLA (2002). *Learning with Kernels*. MIT Press, Cambridge, MA.