THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Optimal Subsampling Designs Under Measurement Constraints

HENRIK IMBERG

Department of Mathematical Sciences

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2023

Optimal Subsampling Designs Under Measurement Constraints
HENRIK IMBERG
ISBN 978-91-7905-826-5

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 5292
ISSN 0346-718X

Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone +46 (0)31-772 1000

Cover:
Illustration of a convex constrained optimisation problem for finding an optimal subsampling design with respect to the L-optimality criterion. Additional details may be found on page 24 in the thesis.

Typeset with LaTeX
Printed by Chalmers digitaltryck
Gothenburg, Sweden 2023

Optimal Subsampling Designs Under Measurement Constraints

HENRIK IMBERG

Department of Mathematical Sciences
Chalmers University of Technology

Abstract

We consider the problem of optimal subsample selection in an experiment setting where observing, or utilising, the full dataset for statistical analysis is practically unfeasible. This may be due to, e.g., computational, economic, or even ethical cost-constraints. As a result, statistical analyses must be restricted to a subset of data. Choosing this subset in a manner that captures as much information as possible is essential.

In this thesis we present a theory and framework for optimal design in general subsampling problems. The methodology is applicable to a wide range of settings and inference problems, including regression modelling, parametric density estimation, and finite population inference. We discuss the use of auxiliary information and sequential optimal design for the implementation of optimal subsampling methods in practice and study the asymptotic properties of the resulting estimators.

The proposed methods are illustrated and evaluated on three problem areas: on subsample selection for optimal prediction in active machine learning (Paper I), optimal control sampling in analysis of safety critical events in naturalistic driving studies (Paper II), and optimal subsampling in a scenario generation context for virtual safety assessment of an advanced driver assistance system (Paper III). In Paper IV we present a unified theory that encompasses and generalises the methods of Paper I–III and introduce a class of expected-distance-minimising designs with good theoretical and practical properties.

In Paper I–III we demonstrate a sample size reduction of 10–50% with the proposed methods compared to simple random sampling and traditional importance sampling methods, for the same level of performance. We propose a novel class of invariant linear optimality criteria, which in Paper IV are shown to reach 90–99% D-efficiency with 90–95% lower computational demand.

Keywords: active sampling, inverse probability weighting, M-estimation, optimal design, unequal probability sampling.

List of publications

This thesis is based on the work contained in the following papers:

I. **Imberg H**, Jonasson J, and Axelson-Fisk M (2020). Optimal sampling in unbiased active learning. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, in *Proceedings of Machine Learning Research*, 108:559–569.

II. **Imberg H**, Lisovskaja V, Selpi, and Nerman O (2022). Optimization of two-phase sampling designs with application to naturalistic driving studies. *IEEE Transactions on Intelligent Transportation Systems*, 23(4):3575–3588.

III. **Imberg H**, Yang X, Flannagan C, and Bärgman J (2022). Active sampling: A machine-learning-assisted framework for finite population inference with optimal subsamples. `arXiv:2212.10024 [stat.ME]`. *Submitted*.

IV. **Imberg H**, Axelson-Fisk M, and Jonasson J (2023). Optimal subsampling designs. `arXiv:2304.03019 [math.ST]`. *Submitted*.

Author contributions

I,II,IV. Derived the theoretical results, developed and implemented the subsampling algorithms, performed the data analysis, drafted and edited the manuscript.

III. Derived the theoretical results, developed and implemented the active sampling algorithm, participated in the interpretation and presentation of results, drafted and edited the manuscript.

Acknowledgements

I would like to express my deepest appreciation to my supervisor Marina Axelson-Fisk for your continuous support and engagement. Thanks for your ceaseless support even when my research ended up in other directions than the ones you envisioned when inviting me to this journey. I also am deeply indebted to my co-supervisor Johan Jonasson, for your perceptive comments on my work, for bringing up the details I have missed, and for always highlighting the importance of mathematical rigour. I owe my thanks to Olle Nerman for introducing me to sampling statistics, and, by doing so, planting the seed of ideas from which this thesis has grown. Thanks to my co-authors, Vera Lisovskaja, Selpi, Xiaomi Yang, Carol Flannagan and Jonas Bärgman. I have really enjoyed our collaboration. Thanks also to Serik Sagitov and Marija Cvijovic for your support along the way.

I would also like to thank all my colleagues and fellow PhD students at the Department of Mathematical Sciences for your company and support during this work. Thanks in particular to Helga Kristín Ólafsdóttir, Oskar Allerbo, Anna Rehammar, and Malin Palö Forsström. Our small talk about family life and statistics has been a frequent source of amusement and a welcome diversion from the duties as a PhD student.

Finally, my greatest gratitude goes to my lovely wife, Ida, and to my wonderful children, Hilma, Aron, and Tyra, for your never-ending love and support. I could not image undertaking this endeavour without you.

Contents

1 Introduction

We consider the problem of optimal subsample selection in an experiment setting where observing, or utilising, the full dataset for statistical analysis is practically unfeasible. This may be due to, e.g., computational, economic, or even ethical cost-constraints. As a result, statistical analyses must be restricted to a subset of data. This problem may be encountered in a wide range of settings and applications, including medical research, bioinformatics, official statistics, machine learning, big data, and traffic safety research. Some specific examples and applications considered in this thesis are described in Chapter 1.1. The general problem is formulated in Chapter 1.2 and the contributions of this thesis summarised in Chapter 1.3.

1.1 Examples and applications

A first example considered in this thesis is the problem of optimal subsample selection for a machine learning and prediction modelling task (Paper I). A common problem in machine learning is the lack of training data available for developing a prediction algorithm. In many cases the inputs to the model are easier to observe than the outcomes, or 'labels'. For instance, electronic sensors and devices make it possible to collect large amounts of data at high speed and low cost. At the same time, some features of the data — often the response variable in the intended prediction model — may require human input in terms of manual annotation, experimentation, or expert judgement. Hence, unlabelled data are abundant whereas labelled data are scarce. Active learning offers a solution to this problem by exploiting information from the observed inputs for selecting which instances to label (MacKay, 1992; Cohn, 1996). This is done in an iterative fashion, alternating between data collection and model fitting, by repeatedly retrieving the labels of new instances. By oversampling the instances that are the most informative with regards to prediction, the

learner may perform better with less data (Lewis and Gale, 1994; Settles, 2012). We study the asymptotic generalisation error of an active learner and derive optimal sampling schemes to minimise the prediction error on unlabelled data. We also use active leaning as a tool to derive optimal subsampling methods assisted by machine learning predictions on yet unseen data, e.g., for estimating a finite population characteristic. Active learning methods are considered further in Chapter 2 and 4.

In Paper II we develop an optimal auxiliary-variable-assisted subsampling method for two-phase sampling studies, with application to naturalistic driving studies. A naturalistic driving study is a study of driving under naturalistic conditions, i.e., without any interventions (Winkelbauer et al., 2010). Data are collected for all driving sessions in a large fleet of vehicles equipped with advanced instrumentation to record vehicle manoeuvres, driver behaviour, and external conditions (van Schagen and Sagberg, 2012). These data are used to describe the characteristics of normal driving, the occurrence of safety critical events (e.g., near crashes and incidents), and to study risk factors for such safety critical events (Dingus et al., 2016). Some of these analyses rely on manual annotation of video sequences, e.g., to extract information on driver behaviour. Since such annotations are extremely time-consuming and expensive, this is usually affordable only for a fraction of the driving sessions in the database. However, a substantial amount of information is already available through automatic recordings of vehicle manoeuvres etc. We demonstrate how such auxiliary information may be utilised to optimise the selection of which instances to annotate with regards to some characteristic of interest, e.g., the detection of potential associations between driving behaviour and a consecutive safety critical event. Optimal auxiliary-variable-assisted subsampling methods are discussed in Chapter 3.

A third application of the methods developed in this thesis is the use of optimal subsampling to reduce computation time in large computer experiments. As an example, in Paper III we consider a method for virtual safety assessment of an advanced driver assistance system. Vehicle safety systems are constantly being developed to improve traffic safety and avoid or mitigate crashes. However, when developing both advanced driver assistance systems and autonomous driving systems, there is a need to assess the impact on safety of the systems before they are on the market. One way to do that is by running virtual simulations and comparing the outcome of simulations both with and without the use of a specific system (Anderson et al., 2013; Seyedi et al., 2021). A drawback with the virtual testing framework is the high computational load. Running all simulations of interest is often too high-dimensional (many simulation parameters varied) to be feasible in practice (Mullins et al., 2018; Sun et al., 2022). Also, even if complete enumeration were feasible, it may not be efficient. A good

estimate may be possible to obtain with much lower computational demand. In this thesis we show how machine learning and sequential optimal design can be utilised to reduce computational load in large-scale computer experiments. These methods will be discussed further in Chapter 4.

As a final example, consider the problem of statistical inference for big data. Massive datasets have become increasingly common in nearly every discipline of science, including health-care (Chen et al., 2017), social science (Righi, 2019), and finance (Óskarsdóttir et al., 2019), to mention a few. The sheer volume of data poses a major challenge for computational and statistical methods for data analysis. A popular approach to handle this issue is by downsizing the dataset through subsampling (Ma et al., 2015; Wang et al., 2018). In Paper IV we develop a general theory of optimal subsampling to ensure that statistical analysis based on a subset of data results in a minimal loss of information. Our contributions are described in Chapter 3 and 4.

1.2 Problem formulation

We consider the problem of estimating a finite population parameter or characteristic on the following form:

i) a (vector) total $\boldsymbol{t_y} = \sum_{i \in \mathcal{D}} \boldsymbol{y}_i$ or function of totals $\boldsymbol{\tau} = \boldsymbol{h}(\boldsymbol{t_y})$, or

ii) a parameter $\boldsymbol{\theta}_0$, defined as the unique solution to an estimation equation

$$\sum_{i \in \mathcal{D}} \boldsymbol{\psi}_i(\boldsymbol{\theta}) = \mathbf{0}, \quad \boldsymbol{\psi}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{v}_i).$$

Here $\boldsymbol{v}_i$ is a data vector associated with an element i in some index set $\mathcal{D}$ of size N, and $\ell(\boldsymbol{\theta}; \boldsymbol{v}_i)$ a loss-function describing the loss associated with the parameter value $\boldsymbol{\theta}$ given data $\boldsymbol{v}_i$. This covers a broad range of inference problems and estimation methods in statistics, maximum likelihood estimation, generalised linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989), quasi-likelihood methods (Wedderburn, 1974), and certain types of M-estimation (Stefanski and Boos, 2002).

We assume that inference using the full-data $\{\boldsymbol{v}_i\}_{i \in \mathcal{D}}$ is unfeasible, e.g., due to practical, computational or economic constraints. Hence, we resort to an approximate solution based on a subset $\mathcal{S}$ of size $n \ll N$. We consider the case when the subset $\mathcal{S}$ is selected by a random mechanism, using unequal probability sampling (Särndal et al., 2003). Thus, each member of the initial

dataset is assigned a strictly positive and possibly unique sampling probability. Using inverse probability weighting, unbiased inference may be assured under very mild assumptions. The inferential process is illustrated in Figure 1.1.

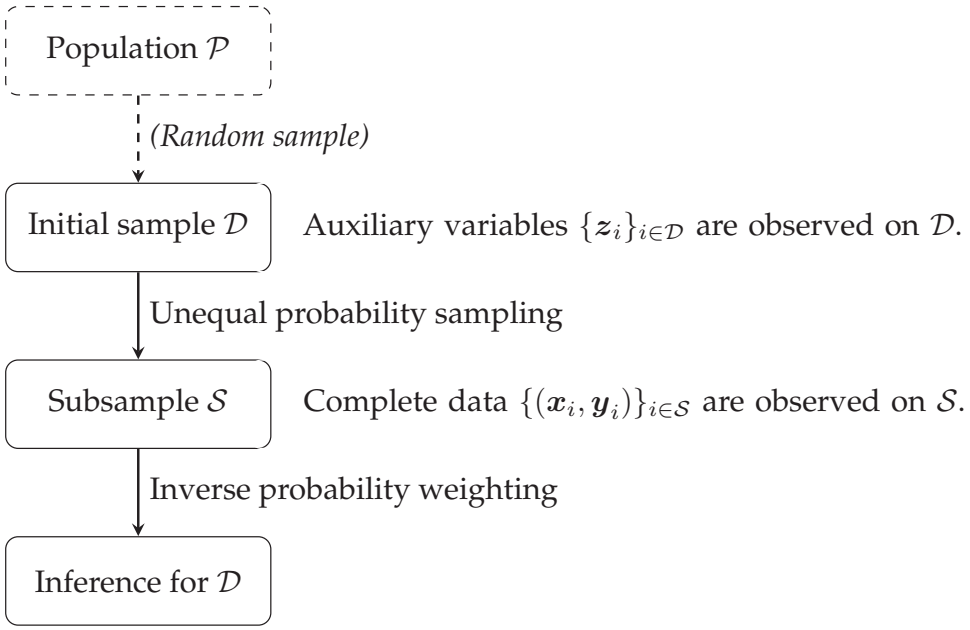


Figure 1.1. Flowchart of the experimental and inferential process in data subsampling applications. In the first step, a sample $\{(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i)\}_{i\in\mathcal{D}}$ is generated and the variables $\{\boldsymbol{z}_i\}_{i\in\mathcal{D}}$ are observed. The study variables of interest, $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, are observed on the subset $\mathcal{S}$, selected using information from the auxiliary variables $\boldsymbol{z}_i$. Estimation is conducted using inverse probability weighting. This ensures that unbiased inference for $\mathcal{D}$ is obtained. Generalisation to an underlying population $\mathcal{P}$ may sometimes also be appropriate. The role of the variables $\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i$ will be made clear in Chapter 2 and 3.

Methods for subdata selection using deterministic procedures have also been proposed, see, e.g., Wynn (1982), Pronzato (2006), Drovandi et al. (2017), Wang et al. (2019), and Deldossi and Tommasi (2022). One major advantage of randomisation-based methods, however, is robustness against model-misspecification. Indeed, the inferential process outlined in Figure 1.1 ensures that unbiased inferences for the characteristics of the initial dataset are obtained under minimal assumptions (Binder, 1983; Pfeffermann, 1993). On the other hand, the increase in variance may be substantial (see, e.g., Korn and Graubard, 1995; Landsman and Graubard, 2013). Hence, the development of optimal subsampling methods is essential to ensure both unbiasedness and high efficiency.

Subsampling problems have been studied for a long time within the field of statistics. Some early and important contributions include the work of

Neyman (1938), Hansen and Hurwitz (1943), and Horvitz and Thompson (1952). Stimulated by modern technological developments, the question of optimal subsampling has attained renewed attention during the past few years. Examples include leverage sampling and approximate numerical linear algebra methods for big data regression (Ma et al., 2015, 2020), optimal subsampling algorithms for binary and multinomial logistic regression (Wang et al., 2018; Yao and Wang, 2019), generalised linear models (Ai et al., 2021b; Zhang et al., 2021; Yu et al., 2022), quantile regression (Ai et al., 2021a; Wang and Ma, 2021), and active learning (Bach, 2007; Kossen et al., 2022; Zhan et al., 2022). However, most of these publications have a highly algorithmic perspective, focusing on a restricted class of models and optimality criteria. Moreover, many of the proposed methods use optimality criteria (e.g., A-optimality) with well-known deficiencies, such as lack of invariance to the measurement-scale of the data and parameterisation of the model. A unified theory of optimal subsampling design is still lacking.

1.3 Contributions

In this thesis we present a theory and framework for optimal design in general subsampling problems. The methodology is applicable to a wide range of problems and settings, including regression modelling, parametric density estimation, and finite population inference. We derive optimality conditions for a broad class of optimality criteria, including A-, D-, E-, and L-optimality. Algorithms to find optimal sampling schemes for both Poisson sampling and multinomial sampling designs are presented. We also study optimal design from an expected-distance-minimising perspective. This naturally leads us to a novel class of linear optimality criteria with good theoretical and practical properties, including computational tractability and invariance under non-singular affine transformations of the data and under a re-parameterisation of the model. We discuss the use of auxiliary information and sequential optimal design for the implementation of optimal subsampling methods in practice and study the asymptotic properties of the resulting estimators. The presented methodology and algorithms are illustrated on problems in machine learning and applications in the traffic safety domain.

The structure of the remainder of this thesis is as follows. In Chapter 2 we provide a brief introduction to survey sampling, active learning, and optimal design. Optimal subsampling designs are discussed in Chapter 3 and methods for sequential optimal design in Chapter 4. A summary of the included papers is provided in Chapter 5. We conclude with a brief discussion and some open problems in Chapter 6.

2 Preliminaries

We start with a brief introduction to survey sampling (Chapter 2.1 and 2.2), optimal design (Chapter 2.3), and active learning (Chapter 2.4). In Chapter 2.1 we consider the estimation of a simple finite population characteristic, such as a total or function of totals. More complex characteristics, such as the minimiser of an empirical risk function or the solution to an estimating equation, are considered in Chapter 2.2.

2.1 Finite population sampling

Consider a finite population, dataset, or index set $\mathcal{D} = \{1, \dots, N\}$ of N elements represented by their indices $i = 1, \dots, N$. Associated with each element $i \in \mathcal{D}$ is a data vector $(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i)$, where $\boldsymbol{x}_i$ is a vector of explanatory variables, covariates, features, or predictors, $\boldsymbol{y}_i$ is a vector of outcomes or response variables, and $\boldsymbol{z}_i$ is a vector of auxiliary variables. Unless otherwise stated, all vectors are assumed to be column vectors. In this section we focus on a finite population characteristic defined by the variables $\boldsymbol{y}_i$. The role of the explanatory and auxiliary variables was briefly described in Chapter 1.2 and will be further discussed in Chapter 2.2 and 3.5, respectively.

We consider a characteristic of the elements in the dataset $\mathcal{D}$ on the following form:

i) a scalar total $t_y = \sum_{i \in \mathcal{D}} y_i$, or corresponding mean t_y/N,

ii) a vector total $\boldsymbol{t_y} = \sum_{i \in \mathcal{D}} \boldsymbol{y}_i$, or corresponding mean vector $\boldsymbol{t_y}/N$, and

iii) a function of totals $\boldsymbol{\tau} = \boldsymbol{h}(\boldsymbol{t_y})$, for a vector of totals $\boldsymbol{t_y} \in \mathbb{R}^q$ and differentiable function $\boldsymbol{h} : \mathbb{R}^q \to \mathbb{R}^p$.

Such a characteristic may be used to describe, e.g., the total energy consumption, unemployment rate, or prevalence of a disease in a population. Examples of characteristics covered by the latter class of statistics (iii) include ratios, simple linear regression coefficients, correlation coefficients, and population variances. More complex characteristics will be considered in Section 2.2.

2.1.1 Unequal probability sampling designs

Now assume that inference based on the full data $\{\boldsymbol{y}_i\}_{i\in\mathcal{D}}$ is unfeasible, e.g., due to practical, economic, or computational constraints. We consider the situation where individual elements $i \in \mathcal{D}$ are selected according to an unequal probability sampling design, i.e., by a random mechanism where each member $i \in \mathcal{D}$ has a strictly positive and possibly unique selection probability (Särndal et al., 2003). Let S_i be the number of times an element $i \in \mathcal{D}$ is selected by the sampling mechanism, and μ_i the corresponding expected number of selections. Sampling may be conducted either with or without replacement. We let $\mathcal{S} = \{i \in \mathcal{D} : S_i > 0\}$ denote the random set of selected elements, and $n = \mathrm{E}[\sum_{i\in\mathcal{D}} S_i]$ the expected size of the subsample.

We assume that sampling is conducted according to one of the following sampling designs:

i) Poisson sampling with replacement (PO-WR): $S_1, \ldots, S_N$ are independent with $S_i \sim \mathrm{Poisson}(\mu_i)$, $\mu_i > 0$.

ii) Poisson sampling without replacement (PO-WOR): $S_1, \ldots, S_N$ are independent with $S_i \sim \mathrm{Bernoulli}(\mu_i)$, $\mu_i \in (0, 1]$.

iii) Multinomial sampling (MULTI): $(S_1, \ldots, S_N) \sim \mathrm{Multinomial}(n, \boldsymbol{\mu}/n)$, $\mu_i \in (0, n)$, $n \in \mathbb{N}$.

For the Poisson sampling designs we note that the independence assumption on $S_1, \ldots, S_N$ implies that the sample size $\sum_{i\in\mathcal{D}} S_i$ is random, with expectation $\mathrm{E}[\sum_{i\in\mathcal{D}} S_i] = \sum_{i\in\mathcal{D}} \mu_i = n$. In contrast, the multinomial design has a fixed sample size. For a given size n, the Poisson and multinomial sampling designs are uniquely determined by the mean vector $\boldsymbol{\mu}$. We say that such a design, for a given size n, is indexed by the sampling scheme $\boldsymbol{\mu}$.

Methods also exist to select a fixed number of elements with unequal probabilities and without replacement, for instance using the conditional Poisson sampling design (Hájek, 1981; Tillé, 2006). Other examples include the Pareto

and Sampford designs, discussed in Grafström (2010). These methods, however, tend to be computationally or analytically intractable, and will therefore not be considered in this thesis. Additional details may be found in, e.g., Tillé (2006), Fuller (2009) and Grafström (2010).

2.1.2 The Hansen-Hurwitz estimator

To account for unequal probabilities of selection, estimation may be performed by sample weighting techniques. For a total $\boldsymbol{t}_{\boldsymbol{y}}$, we consider the Hansen-Hurwitz estimator (Hansen and Hurwitz, 1943)

$$\hat{\boldsymbol{t}}_{\boldsymbol{y}} = \sum_{i \in \mathcal{S}} S_i w_i \boldsymbol{y}_i, \quad w_i = 1/\mu_i. \tag{2.1}$$

When sampling is without replacement, this estimator coincides with the Horvitz-Thompson estimator for $\boldsymbol{t}_{\boldsymbol{y}}$ (Horvitz and Thompson, 1952). Moreover, (2.1) is an unbiased estimator of the total $\boldsymbol{t}_{\boldsymbol{y}}$, provided that $\mu_i > 0$ for all $i \in \mathcal{D}$ (see, e.g., Särndal et al., 2003). A corresponding estimator for the mean $\boldsymbol{t}_{\boldsymbol{y}}/N$ is given by $\hat{\boldsymbol{t}}_{\boldsymbol{y}}/N$ or $\hat{\boldsymbol{t}}_{\boldsymbol{y}}/\hat{N}$, where $\hat{N} = \sum_{i \in \mathcal{S}} S_i w_i$, and an estimator for a function of totals $\boldsymbol{\tau} = \boldsymbol{h}(\boldsymbol{t}_{\boldsymbol{y}})$ by $\hat{\boldsymbol{\tau}} = \boldsymbol{h}(\hat{\boldsymbol{t}}_{\boldsymbol{y}})$.

Since $\hat{\boldsymbol{t}}_{\boldsymbol{y}}$ is linear in the random variables S_i, the covariance matrix of the estimator $\hat{\boldsymbol{t}}_{\boldsymbol{y}}$ is on the form

$$\mathbf{Cov}(\hat{\boldsymbol{t}}_{\boldsymbol{y}}) = \sum_{i,j \in \mathcal{D}} \frac{\mathrm{Cov}(S_i, S_j)}{\mu_i \mu_j} \boldsymbol{y}_i \boldsymbol{y}_i^{\mathsf{T}} \tag{2.2}$$

$$= \begin{cases} \sum_{i \in \mathcal{D}} w_i \boldsymbol{y}_i \boldsymbol{y}_i^{\mathsf{T}} & \text{for PO-WR,} \\ \sum_{i \in \mathcal{D}} w_i \boldsymbol{y}_i \boldsymbol{y}_i^{\mathsf{T}} - \boldsymbol{t}_{\boldsymbol{y}\boldsymbol{y}^{\mathsf{T}}} & \text{for PO-WOR,} \\ \sum_{i \in \mathcal{D}} w_i \boldsymbol{y}_i \boldsymbol{y}_i^{\mathsf{T}} - \boldsymbol{t}_{\boldsymbol{y}} \boldsymbol{t}_{\boldsymbol{y}}^{\mathsf{T}}/n & \text{for MULTI,} \end{cases} \tag{2.3}$$

where $\boldsymbol{t}_{\boldsymbol{y}\boldsymbol{y}^{\mathsf{T}}} = \sum_{i \in \mathcal{D}} \boldsymbol{y}_i \boldsymbol{y}_i^{\mathsf{T}}$ (Tillé, 2006). Corresponding results for the estimator $\hat{\boldsymbol{\tau}} = \boldsymbol{h}(\hat{\boldsymbol{t}}_{\boldsymbol{y}})$ of a function of totals are provided below.

Under suitable assumptions it holds that

$$\sqrt{n}(\hat{\boldsymbol{t}}_{\boldsymbol{y}} - \boldsymbol{t}_{\boldsymbol{y}}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_0)$$

as $n \to \infty$, $N \to \infty$ and $n/N \to \gamma \in (0, 1)$, where $\boldsymbol{\Gamma}_0$ is the limiting value of $n\mathbf{Cov}(\hat{\boldsymbol{t}}_{\boldsymbol{y}})$ (Fuller, 2009). The convergence in distribution is under the law of the sampling design, and we say that the Hansen-Hurwitz estimator is design-consistent for $\boldsymbol{t}_{\boldsymbol{y}}$ (Särndal et al., 2003). By the delta method (Cramér, 1946; Sen

and Singer, 1993) it follows that for a differentiable function $\boldsymbol{h}(\boldsymbol{u})$, the estimator $\hat{\boldsymbol{\tau}} = \boldsymbol{h}(\hat{\boldsymbol{t}}_{\boldsymbol{y}})$ is also approximately Gaussian in sufficiently large samples, with mean

$$\mathrm{E}[\hat{\boldsymbol{\tau}}] = \boldsymbol{\tau} + o(n^{-1/2})$$

and covariance matrix

$$\mathbf{Cov}(\hat{\boldsymbol{\tau}}) = \mathbf{J}_{\boldsymbol{h}}(\boldsymbol{t}_{\boldsymbol{y}})\mathbf{Cov}(\hat{\boldsymbol{t}}_{\boldsymbol{y}})\mathbf{J}_{\boldsymbol{h}}(\boldsymbol{t}_{\boldsymbol{y}})^{\mathsf{T}} + o(n^{-1}).$$

Here $o(n^{-1/2})$ and $o(n^{-1})$ denote elementwise convergence to zero at rate $n^{-1/2}$ and n^{-1}, respectively. $\mathbf{J}_{\boldsymbol{h}}(\boldsymbol{u})$ is the Jacobian matrix of $\boldsymbol{h}(\boldsymbol{u})$, i.e., the matrix with rows $\nabla h_j(\boldsymbol{u})^{\mathsf{T}}$.

2.1.3 Variance estimation

For variance estimation, note that the covariance matrix (2.2) itself is a (matrix-valued) finite population total. Hence, variance estimation may be conducted by similar sample-weighting techniques as described above. For fixed-size designs, however, an alternative estimator due to Sen (1953) and Yates and Grundy (1953) is often advocated. Specifically, the following unbiased covariance matrix estimators are commonly employed:

$$\widehat{\mathbf{Cov}}(\hat{\boldsymbol{t}}_{\boldsymbol{y}}) = \begin{cases} \sum_{i\in\mathcal{S}} S_i w_i^2 \boldsymbol{y}_i \boldsymbol{y}_i^{\mathsf{T}} & \text{for PO-WR,} \\ \sum_{i\in\mathcal{S}} S_i w_i (w_i - 1) \boldsymbol{y}_i \boldsymbol{y}_i^{\mathsf{T}} & \text{for PO-WOR,} \\ \frac{n}{n-1} \sum_{i\in\mathcal{S}} S_i \left(w_i \boldsymbol{y}_i - \hat{\boldsymbol{t}}_{\boldsymbol{y}}/n\right)\left(w_i \boldsymbol{y}_i - \hat{\boldsymbol{t}}_{\boldsymbol{y}}/n\right)^{\mathsf{T}} & \text{for MULTI.} \end{cases}$$

An estimator for the covariance matrix of $\hat{\boldsymbol{\tau}} = \boldsymbol{h}(\hat{\boldsymbol{t}}_{\boldsymbol{y}})$ is then obtained as

$$\widehat{\mathbf{Cov}}(\hat{\boldsymbol{\tau}}) = \mathbf{J}_{\boldsymbol{h}}(\hat{\boldsymbol{t}}_{\boldsymbol{y}})\widehat{\mathbf{Cov}}(\hat{\boldsymbol{t}}_{\boldsymbol{y}})\mathbf{J}_{\boldsymbol{h}}(\hat{\boldsymbol{t}}_{\boldsymbol{y}})^{\mathsf{T}}.$$

See, e.g., Särndal et al. (2003), Tillé (2006), and Fuller (2009) for further details.

2.2 Empirical risk minimisation

Now consider a p-dimensional parameter $\boldsymbol{\theta}_0$ defined by

$$\boldsymbol{\theta}_0 = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\boldsymbol{\Omega}} \ell_0(\boldsymbol{\theta}), \tag{2.4}$$

i.e., as the minimiser of some function $\ell_0(\boldsymbol{\theta})$ over some parameter space $\boldsymbol{\Omega} \subset \mathbb{R}^p$. We assume further that $\boldsymbol{\theta}_0$ is unique, and that $\ell_0(\boldsymbol{\theta})$ is twice differentiable and

can be written on the form

$$\ell_0(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}), \quad \ell_i(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}; \boldsymbol{v}_i). \tag{2.5}$$

The individual contributions $\ell_i(\boldsymbol{\theta})$ are assumed to be functions of the parameter $\boldsymbol{\theta}$ and data vectors $\boldsymbol{v}_i$. We recognise (2.4) as an empirical risk minimisation problem (Vapnik, 1991), and hence refer to $\boldsymbol{\theta}_0$ as the (full-data) empirical risk minimiser (ERM).

Under the above assumptions, $\boldsymbol{\theta}_0$ may also be defined as the unique solution to the estimation equation

$$\sum_{i \in \mathcal{D}} \boldsymbol{\psi}_i(\boldsymbol{\theta}) = \mathbf{0}, \quad \boldsymbol{\psi}_i(\boldsymbol{\theta}) = \nabla \ell_i(\boldsymbol{\theta}), \tag{2.6}$$

with $\mathcal{D} = \{1, \ldots, N\}$. This setting covers a broad range of inference problems, models, and estimation methods in statistics, including maximum likelihood estimation, generalised linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989), quasi-likelihood methods (Wedderburn, 1974), and certain types of M-estimation (Stefanski and Boos, 2002). Some specific examples include:

i) Inference for a finite population mean: consider a finite population of N individuals, where each individual is associated with a non-random vector characteristic $\boldsymbol{y}_i$. The vector of finite population means $N^{-1} \sum_{i=1}^{N} \boldsymbol{y}_i$ may be written on the form (2.4)–(2.6) with $\boldsymbol{v}_i = \boldsymbol{y}_i$ and $\ell(\boldsymbol{\theta}; \boldsymbol{v}_i) = ||\boldsymbol{y}_i - \boldsymbol{\theta}||_2^2 = (\boldsymbol{y}_i - \boldsymbol{\theta})^\mathsf{T}(\boldsymbol{y}_i - \boldsymbol{\theta})$.

ii) Parametric density estimation: given independent and identically distributed data $y_1, \ldots, y_N$ from a probability distribution with density function $f_{\boldsymbol{\theta}}(y)$, the maximum likelihood estimate of $\boldsymbol{\theta}$ may be written on the form (2.4)–(2.6) with $\boldsymbol{v}_i = y_i$ and $\ell(\boldsymbol{\theta}; \boldsymbol{v}_i) = -\log f_{\boldsymbol{\theta}}(y_i)$.

iii) Regression modelling: consider a random sample $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, a vector of regression coefficients $\boldsymbol{\theta}$, a (non-linear) model $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ for the conditional mean of Y given $\boldsymbol{x}$, and a differentiable loss-function $l : \mathbb{R}^2 \to \mathbb{R}_+$ such that $l(\hat{y}, y) = 0$ if and only if $\hat{y} = y$. With $\boldsymbol{v}_i = (\boldsymbol{x}_i, y_i)$ and $\ell(\boldsymbol{\theta}; \boldsymbol{v}_i) = l(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y_i)$, the equations (2.4)–(2.6) define an estimate of the vector of regression coefficients $\boldsymbol{\theta}$.

2.2.1 The Hansen-Hurwitz ERM

Assume, as before, that complete data $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ can only be observed for a subset $\mathcal{S} \subset \mathcal{D}$, which is selected using unequal probability sampling. We consider an estimator for $\boldsymbol{\theta}_0$ on the form

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}} &= \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \boldsymbol{\Omega}} \hat{\ell}_{\boldsymbol{\mu}}(\boldsymbol{\theta}), \\ \hat{\ell}_{\boldsymbol{\mu}}(\boldsymbol{\theta}) &= \sum_{i \in \mathcal{S}} S_i w_i \ell_i(\boldsymbol{\theta}), \quad w_i = 1/\mu_i, \end{aligned} \tag{2.7}$$

where S_i is the number of times an element $i \in \mathcal{D}$ is selected by the sampling mechanism, μ_i is the corresponding expected number of selections, and $\mathcal{S} = \{i \in \mathcal{D} : S_i > 0\}$ is the random set of selected elements. We recognise this as the Hansen-Hurwitz estimator of the full-data empirical risk function (2.5). Hence, we refer to $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ as the Hansen-Hurwitz empirical risk minimiser.

It has been shown by Binder (1983) that under suitable regularity conditions the distribution of the estimator (2.7) with respect to the sampling mechanism is approximately Gaussian with mean

$$\mathrm{E}[\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}] = \boldsymbol{\theta}_0 + o(n^{-1/2}),$$

and covariance matrix

$$\begin{aligned} \mathbf{Cov}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}} - \boldsymbol{\theta}_0) &= \boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0) + o(n^{-1}), \\ \boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0) &= \mathbf{H}(\boldsymbol{\theta}_0)^{-1} \mathbf{V}(\boldsymbol{\mu}; \boldsymbol{\theta}_0) \mathbf{H}(\boldsymbol{\theta}_0)^{-1}. \end{aligned}$$

Here $\mathbf{H}(\boldsymbol{\theta}_0) = \frac{\partial^2 \ell_0(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\mathsf{T}}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ is the Hessian of the full-data empirical risk function (2.5) at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$,

$$\mathbf{V}(\boldsymbol{\mu}; \boldsymbol{\theta}_0) = \mathbf{Cov}\left(\nabla \hat{\ell}_{\boldsymbol{\mu}}(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right) = \sum_{i,j \in \mathcal{D}} \frac{\mathrm{Cov}(S_i, S_j)}{\mu_i \mu_j} \boldsymbol{\psi}_i(\boldsymbol{\theta}_0) \boldsymbol{\psi}_j(\boldsymbol{\theta}_0)^\mathsf{T}$$

the covariance matrix of the gradient $\nabla \hat{\ell}_{\boldsymbol{\mu}}(\boldsymbol{\theta})$ with respect to the sampling mechanism, evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, and $\boldsymbol{\psi}_i(\boldsymbol{\theta}) = \nabla \ell_i(\boldsymbol{\theta})$. For further details we refer to Binder (1983) and Fuller (2009).

It follows from (2.2)–(2.3) and (2.6) that the matrix $\mathbf{V}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$ can be simplified by

$$\mathbf{V}(\boldsymbol{\mu}; \boldsymbol{\theta}_0) = \begin{cases} \sum_{i \in \mathcal{D}} w_i \boldsymbol{\psi}_i(\boldsymbol{\theta}_0) \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T} & \text{for PO-WR and MULTI,} \\ \sum_{i \in \mathcal{D}} (w_i - 1) \boldsymbol{\psi}_i(\boldsymbol{\theta}_0) \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T} & \text{for PO-WOR.} \end{cases}$$

2.2.2 Variance estimation

Variance estimation may be conducted by using the plug-in estimator

$$\widehat{\boldsymbol{\Gamma}}(\boldsymbol{\mu};\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}) = \widehat{\mathbf{H}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}})^{-1}\widehat{\mathbf{V}}(\boldsymbol{\mu};\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}})\widehat{\mathbf{H}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}})^{-1},$$

where

$$\widehat{\mathbf{H}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}) = \sum_{i\in\mathcal{S}} S_i w_i \frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}}$$

is an estimator for the full data Hessian $\mathbf{H}(\boldsymbol{\theta}_0)$, and

$$\widehat{\mathbf{V}}(\boldsymbol{\mu};\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}) = \begin{cases} \sum_{i\in\mathcal{S}} S_i w_i^2 \boldsymbol{\psi}_i(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}})\boldsymbol{\psi}_i(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}})^{\mathsf{T}} & \text{for PO-WR and MULTI,} \\ \sum_{i\in\mathcal{S}} S_i w_i(w_i-1) \boldsymbol{\psi}_i(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}})\boldsymbol{\psi}_i(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}})^{\mathsf{T}} & \text{for PO-WOR,} \end{cases}$$

is an estimator for $\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$. A formal justification may be found in Binder (1983).

2.2.3 Design-based inference and super-population inference

The main view on inference adopted in this thesis is a design-based perspective where the study variables are considered as fixed but unknown constants. Conditioned on the initial dataset $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i\in\mathcal{D}}$, all randomness involved in the subsampling experiment is due to the sample selection mechanism. The statistical properties of the resulting estimator are consequently formulated under the law of the design, which is under immediate control of the investigator. An important consequence is that, using the sample weighting techniques described above, the validity of the inference made from a probability sample $\mathcal{S}$ about the characteristics of the full data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i\in\mathcal{D}}$, is free of model assumptions. Hence, a design-consistent estimator is obtained even under the realistic assumption of model-misspecification, i.e., when the assumed model does not hold true in the underlying population (Skinner, 1989; Pfeffermann, 1993). This is an important robustness property of design-based estimation. Model-based methods, on the other hand, produce more efficient estimators when the assumptions of the model are fulfilled, but are more sensitive to model-misspecification (Cramér, 1946; Korn and Graubard, 1995; Shimodaira, 2000). In this thesis we pursue a design-based approach to ensure unbiasedness and develop optimal subsampling methods to increase efficiency.

In many applications the scope of inference goes beyond that of the initial dataset $\mathcal{D}$, and one may wish to generalise the findings to an underlying population $\mathcal{P}$ (Figure 1.1, Chapter 1). In the survey sampling literature, this

is commonly referred to as super-population inference (Hartley and Sielken, 1975). Indeed, we may often view our estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ as an estimator for an underlying super-population parameter $\boldsymbol{\theta}^*$, defined as the limiting value of $\boldsymbol{\theta}_0$ as the size N of the initial sample $\mathcal{D}$ tends to infinity, and with the data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ being generated according to its limiting empirical distribution. Rubin-Bleuer and Kratina (2005) have established asymptotic normality and consistency of $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ as an estimator for the super-population parameter $\boldsymbol{\theta}^*$ under suitable conditions. By the law of total covariance, the covariance matrix of $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ around $\boldsymbol{\theta}^*$ is on the form

$$\mathbf{Cov}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}} - \boldsymbol{\theta}^*) = \mathbf{Cov}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*) + \mathrm{E}\left[\boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)\right] + o(n^{-1}).$$

The first term denotes the covariance matrix of the full-data parameter $\boldsymbol{\theta}_0$ around the super-population parameter $\boldsymbol{\theta}^*$. The second term is the expectation of the covariance matrix of $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ around $\boldsymbol{\theta}_0$. For further details we refer to Rubin-Bleuer and Kratina (2005) and Fuller (2009).

2.3 Optimal design

The overall aim of this thesis is to develop optimal subsampling methods. To do so we need a theory of optimal design. Therefore, consider a class of experiments Ξ and corresponding consistent estimators $\hat{\boldsymbol{\theta}}_\xi, \xi \in \Xi$ for an unknown parameter $\boldsymbol{\theta}^*$, with unequal covariance matrices $\boldsymbol{\Gamma}_\xi$. In a subsampling application, the experiment is determined by the choice of sampling design and sampling scheme. Ideally, we would like to find an experiment $\xi^* \in \Xi$ such that $\boldsymbol{\Gamma}_\xi - \boldsymbol{\Gamma}_{\xi^*}$ is positive semi-definite for all $\xi \in \Xi$. Such universal optimality, however, is not possible to achieve in general. Hence, instead we consider a function $\Phi : \boldsymbol{S}_+^{p\times p} \to \mathbb{R}$ on the set of real, symmetric, positive semi-definite $p \times p$ matrices, for which a minimiser $\xi^* \in \Xi$ is sought. For Φ to be a meaningful measure of optimality we require the function to be monotone for Loewner's ordering, i.e., that

$$\Phi(\mathbf{U}) \geq \Phi(\mathbf{V}) \text{ for all } \mathbf{U}, \mathbf{V} \in \boldsymbol{S}_+^{p\times p} \text{ such that } \mathbf{U} \geq \mathbf{V}, \tag{2.8}$$

with $\mathbf{U} \geq \mathbf{V}$ meaning that $\mathbf{U} - \mathbf{V}$ is positive semi-definite (Pukelsheim, 1993).

2.3.1 Criteria of optimality

Some popular optimality criteria are defined and summarised in Table 2.1. These include the D-optimality criterion (minimise the determinant of the

covariance matrix), E-optimality criterion (minimise the largest eigenvalue of the covariance matrix) and L-optimality criterion (minimise the average variance of a collection of linear combinations $\mathbf{L}^\mathsf{T}\hat{\boldsymbol{\theta}}_\xi$). Two important special cases of the L-optimality criterion are the A-optimality criterion (minimise the average variance) and c-optimality criterion (minimise the variance of a linear combination $\mathbf{c}^\mathsf{T}\hat{\boldsymbol{\theta}}_\xi$), obtained with $\mathbf{L} = \mathbf{I}_{p\times p}$ and $\mathbf{L} = \mathbf{c}$, respectively, where $\mathbf{I}_{p\times p}$ is the $p \times p$ identity matrix and $\mathbf{c}$ a non-zero $p \times 1$ vector (Silvey, 1980; Atkinson and Donev, 1992).

Table 2.1. Definition of the A-, c-, D-, E-, L-, and V-optimality criteria. Φ is a real-valued function on the set of real symmetric positive semi-definite $p \times p$ matrices, $\boldsymbol{\Gamma}$ the $p \times p$ covariance matrix of an estimator $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$, $\lambda_{\max}(\boldsymbol{\Gamma})$ the largest eigenvalue of $\boldsymbol{\Gamma}$, $\mathbf{c}$ a non-zero $p \times 1$ vector, and $\mathbf{L}$ a non-zero $p \times m$ matrix with columns $\mathbf{a}_1, \ldots, \mathbf{a}_m$. $\mathcal{X}$ is the set of possible values for the predictors $\boldsymbol{x}$ and $\boldsymbol{\varphi} : \mathcal{X} \to \mathbb{R}^p$ a feature expansion of the data.

Optimality criterion	Description	Objective function $\Phi(\boldsymbol{\Gamma})$
A-optimality	Minimise average variance, minimise trace of covariance matrix, minimise sum of eigenvalues.	$\sum_{i=1}^{p} \operatorname{Var}(\hat{\theta}_i) = \operatorname{tr}(\boldsymbol{\Gamma})$
c-optimality	Minimise variance of a linear combination or contrast $\mathbf{c}^\mathsf{T}\hat{\boldsymbol{\theta}}$.	$\operatorname{Var}(\mathbf{c}^\mathsf{T}\hat{\boldsymbol{\theta}}) = \mathbf{c}^\mathsf{T}\boldsymbol{\Gamma}\mathbf{c}$
D-optimality	Minimise generalised variance, minimise determinant, minimise product of eigenvalues.	$\det(\boldsymbol{\Gamma})^{1/p}$ or $\log\det(\boldsymbol{\Gamma})$
E-optimality	Minimise maximal eigenvalue, minimise variance along the direction of largest uncertainty.	$\lambda_{\max}(\boldsymbol{\Gamma})$
L-optimality	Minimise average variance of a collection of linear combinations or contrasts $\mathbf{L}^\mathsf{T}\hat{\boldsymbol{\theta}}$.	$\sum_{i=1}^{m} \operatorname{Var}(\mathbf{a}_i^\mathsf{T}\hat{\boldsymbol{\theta}}) = \operatorname{tr}(\boldsymbol{\Gamma}\mathbf{L}\mathbf{L}^\mathsf{T})$
V-optimality	Minimise average prediction variance with respect to measure $\nu(\boldsymbol{x})$ on $\mathcal{X}$, assuming a linear model $\hat{y} = \boldsymbol{\varphi}(\boldsymbol{x})^\mathsf{T}\hat{\boldsymbol{\theta}}$.	$\int_{\mathcal{X}} \operatorname{Var}(\boldsymbol{\varphi}(\boldsymbol{x})^\mathsf{T}\hat{\boldsymbol{\theta}}) d\nu(\boldsymbol{x}) = \operatorname{tr}\left(\boldsymbol{\Gamma}\int_{\mathcal{X}} \boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x})^\mathsf{T} d\nu(\boldsymbol{x})\right)$

The A-, D- and E- optimality criteria have a simple geometric interpretation as follows. Consider the random set $\mathcal{C}(\hat{\boldsymbol{\theta}}_\xi) := \{\boldsymbol{\theta} \in \mathbb{R}^p : (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_\xi)^\mathsf{T}\boldsymbol{\Gamma}_\xi^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_\xi) \leq \chi^2_{p,\alpha}\}$, where $\chi^2_{p,\alpha}$ is the α-quantile of a χ^2-distribution with p degrees of freedom. For an (approximately) normally distributed estimator $\hat{\boldsymbol{\theta}}_\xi$, this defines an (approximate) $100 \times (1-\alpha)\%$ ellipsoidal confidence set for $\boldsymbol{\theta}^*$ in $\mathbb{R}^p$. D-optimality minimises the volume of this confidence ellipsoid over the class of experiments Ξ. E-optimality minimises the length of its longest axis, and A-optimality the length of the diagonal of the minimal bounding box (parallelotope) around the confidence ellipsoid (Figure 2.1) (Pronzato and Pázman, 2013).

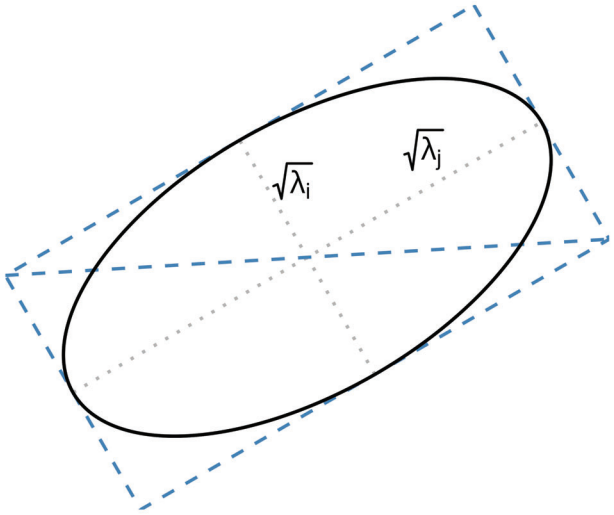


Figure 2.1. Confidence ellipse for a parameter $\boldsymbol{\theta}^* \in \mathbb{R}^2$. The ellipse is centred at the estimated value of $\boldsymbol{\theta}^*$. The D-optimality criterion minimises the area (volume) of the confidence ellipsoid, which is proportional to the square root of the product of the eigenvalues of $\boldsymbol{\Gamma}_\xi$, over all experiments $\xi \in \Xi$. The E-optimality criterion minimises its longest axis, i.e., the largest eigenvalue of $\boldsymbol{\Gamma}_\xi$. The A-optimality criterion minimises the length of the diagonal of the minimal bounding box around the confidence ellipsoid, which is equivalent to minimising the sum of the eigenvalues of $\boldsymbol{\Gamma}_\xi$. Illustration adopted from Geuten et al. (2007).

Another popular optimality criterion is the V-optimality criterion, which minimises the average prediction variance with respect to some measure $\nu(\boldsymbol{x})$ on the design space $\mathcal{X}$ (Welch, 1984). As it turns out, this is a linear optimality criterion and hence is covered by the L-optimality criterion for a matrix $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}^\mathsf{T} = \int_\mathcal{X} \boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x})^\mathsf{T} d\nu(\boldsymbol{x})$ (Table 2.1) (Atkinson and Donev, 1992). A natural choice for the measure $\nu(\boldsymbol{x})$ in data subsampling problems is the empirical measure on $\{\boldsymbol{x}_i\}_{i\in\mathcal{D}}$.

A property that is often desirable of an optimal design, is invariance under a non-singular affine transformation of the data and under a re-parameterisation of the model. That is, the optimal design and the statistical properties of the resulting estimator should not depend on the choice of parameterisation, nor on the scaling or coding of the data prior to modelling. The most common example of a transformation- and parameterisation-invariant optimality criterion is the D-optimality criterion. In contrast, the A- and E-optimality criteria are sensitive to changes in the parameterisation or data, and hence lack such invariance properties (Atkinson and Donev, 1992). An L-optimal design may or may not be transformation- and parameterisation-invariant, depending on whether or not the coefficient matrix $\mathbf{L}$ of the L-optimality criterion is adapted to the

parameterisation of the problem and scaling of the data. Some invariant L-optimal designs are discussed in Chapter 3.4.

2.3.2 Linear optimality criteria

Linear optimality criteria play a central role in this thesis. For instance, in Paper I we consider an optimal sampling scheme to minimise the mean squared prediction error and asymptotic generalisation error of an unbiased active learning algorithm. The result is obtained by establishing equivalence to the L-optimality criterion. In Paper II we consider various linear optimality criteria when studying risk factors for an adverse event in a case-control setting. The c-optimality criterion is used in Paper III to derive an optimal subsampling method for estimating a finite population characteristic.

One of the main reasons for considering linear optimality criteria is due to its analytical and computational tractability for optimal design in data subsampling problems. The L-optimality criterion corresponds to a convex optimisation problem for which a simple closed-form solution exists (see Chapter 3.2). Consequently, in Paper IV we demonstrate a reduction in computation time by more than 90% for finding an optimal sampling scheme for the L-optimality criterion compared to non-linear criteria. The L-optimality criterion also plays a central role in the machinery for finding optimal sampling schemes for more complex, non-linear optimality criteria (see Chapter 3.3).

A crucial issue with the L-optimality criterion is the choice of coefficient matrix $\mathbf{L}$, i.e., the linear combinations for which an optimal design is sought. In Paper I–III we discuss different options depending on the aim of the study. In Paper IV we introduce a novel class of linear optimality criteria derived by minimising the expected distance of our estimator from the target characteristic, with respect to some suitable statistical distance function. For instance, we consider the Kullback-Leibler divergence, empirical risk distance, and Mahalanobis distance. The resulting optimal designs are shown to have good invariance properties, low computational complexities, and high D-efficiencies. This class of optimality criteria will be considered further in Chapter 3.4.

2.3.3 Sequential optimal design

Beyond simple linear models, the covariance matrices $\mathbf{\Gamma}_\xi$ and the corresponding optimal design ξ^* generally depend on the true value of the parameter $\boldsymbol{\theta}$, which in practice is unknown (Pronzato, 2006). Sequential optimal design of-

fers a solution to this problem by iteratively adding new design points, at each step using the information available for choosing an experiment for the next step (Box and Hunter, 1965; Ford and Silvey, 1980). An example of a sequential optimal design method in a machine learning context is active learning, which is discussed next. We will return to sequential optimal design methods for data subsampling in Chapter 4.

2.4 Active learning

Now consider the problem of predicting an outcome Y given some input $\boldsymbol{x}$. The outcome may either be categorical, as in classification problems, or numeric, as in regression problems. Assume further that we are given a large collection $\{\boldsymbol{x}_i\}_{i=1}^N$ of such inputs, but that the corresponding outcomes $\{y_i\}_{i=1}^N$ can be observed only for a smaller subset $\mathcal{S}$ of size $n \ll N$. This is the problem setting in active learning, a branch of machine learning dealing with the optimal subdata selection problem in a prediction modelling context. Active learning algorithms differ from traditional 'passive' learning methods in that the machine learning algorithm itself chooses the data from which it learns. This is done by exploring and exploiting structures in the data to enhance learning, and enables machine learning models to perform better with less training (Settles, 2012).

The active learning process is illustrated in Figure 2.2 and 2.3. The algorithm starts with small initial sample, usually a simple random sample, and iteratively queries the labels y_i of yet unlabelled instances. This is typically done by asking a human annotator or expert for feedback regarding the value of y_i. As more labelled data becomes available, the model is re-trained and new instances are selected to maximise the predictive ability of the model. Some popular querying strategies include uncertainty sampling (Lewis and Catlett, 1994), hypothesis reduction techniques (Haussler, 1989), and variance reduction techniques (MacKay, 1992; Cohn, 1996; Schein and Ungar, 2007). The first class of methods proceeds by querying the label of the instances that the prediction algorithm currently is most uncertain of. Hypothesis reduction techniques, on the other hand, aim to reduce the version space as much as possible, i.e., the set of hypotheses or prediction rules that are compatible with observed data. Variance reduction techniques utilise methods from optimal design to minimise the prediction variance.

A potential drawback of active learning methods is the sensitivity to model misspecification. Active learning causes a change in the distribution of the features $\boldsymbol{x}$ in the labelled training dataset compared to the distribution of the

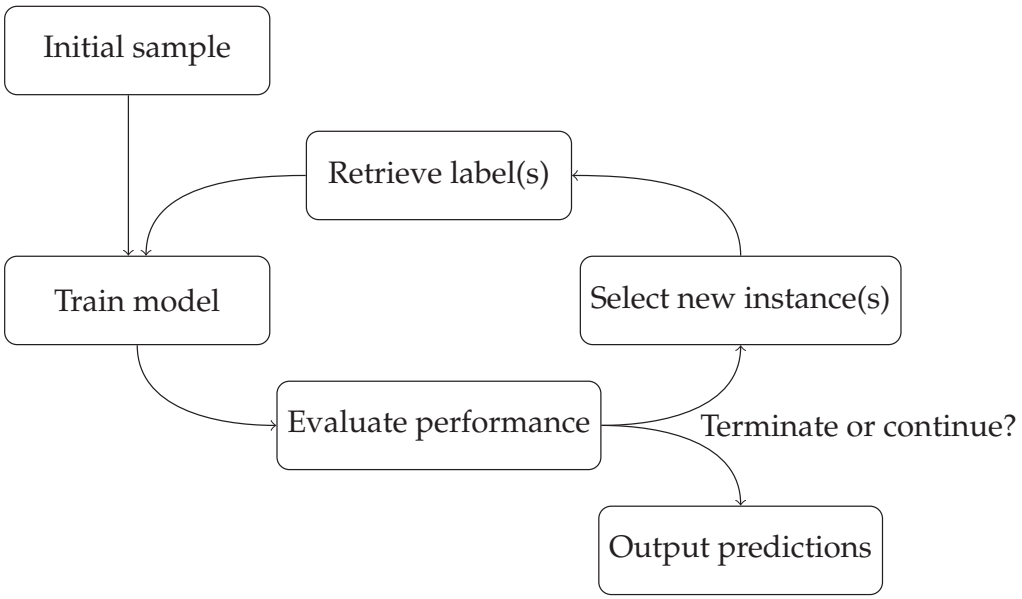


Figure 2.2. Flowchart of the active learning process. The algorithm starts with a small initial sample and iteratively retrieves new training data, whereafter the model is updated accordingly. The algorithm is terminated when a fixed number of instances have been selected, or some stopping criterion on the performance of the model reached.

unlabelled data. This is commonly referred to as covariate shift (Shimodaira, 2000; Quinonero-Candela et al., 2008). For a discriminative model, i.e., a model for the conditional distribution of the outcome Y given covariates $\boldsymbol{x}$, covariate shift can be safely ignored when the model is correctly specified. Indeed, when the model is conditioned on the features causing the covariate shift, the selection mechanism is ignorable (Rubin, 1976; Pfeffermann and Sverchkov, 1999; Zadrozny, 2004). For a misspecified model, however, this is not longer the case and a covariate shift may result in inconsistent estimates and poor predictions (Shimodaira, 2000). Various solutions to this problem have been proposed, including weighting the loss function (Shimodaira, 2000; Sugiyama, 2006; Sugiyama and Nakajima, 2009), robust experiment design under model-misspecification (Tommasi, 2012; Meng et al., 2021; Sirpitzi et al., 2023), and design-based unbiased active learning methods (Bach, 2007; Beygelzimer et al., 2009; Chu et al., 2011; Ganti and Gray, 2012).

Active learning is considered further in two of the papers included in this thesis. In Paper I we consider the active learning problem from a finite population sampling perspective. We derive optimal sampling schemes to minimise the mean squared prediction error and asymptotic generalisation error of an unbiased active learning algorithm. In Paper III we explore the possibilities of using active learning also in applications where prediction is not of primary

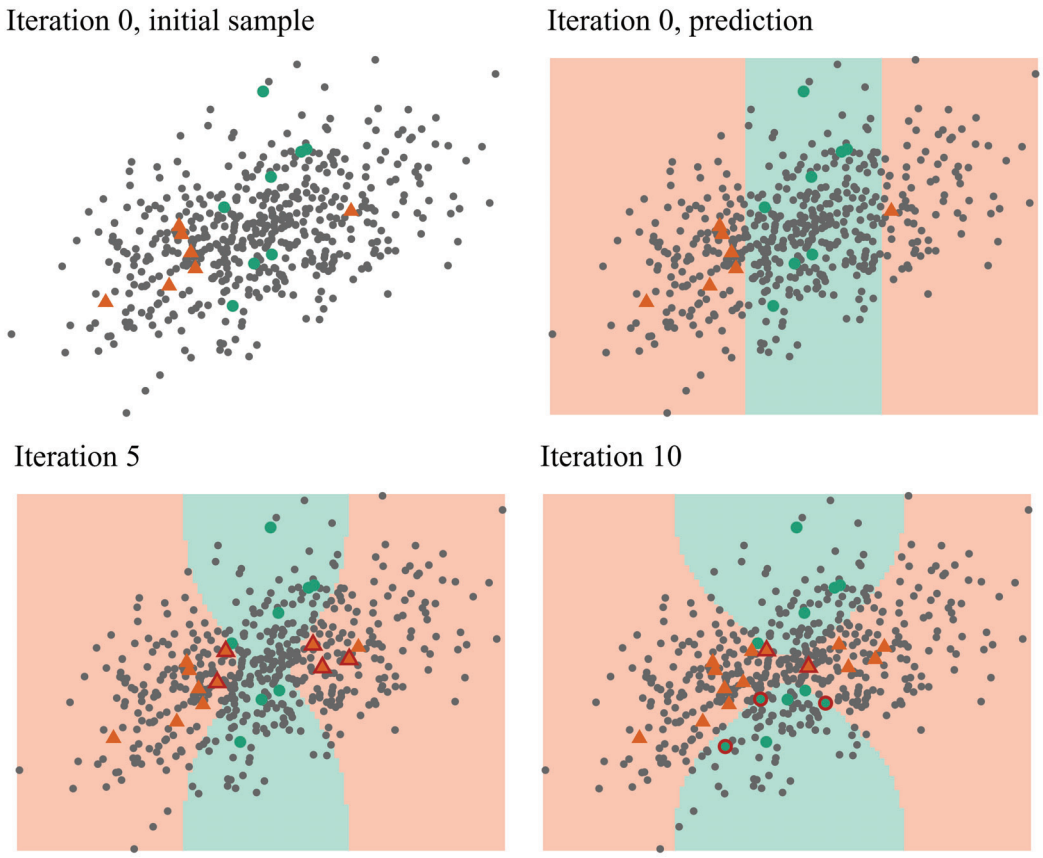


Figure 2.3. Illustration of the active learning process for a binary classification problem. The algorithm starts with a small initial (simple random) sample of labelled data points. New training examples are selected and labelled iteratively, and the prediction model is updated accordingly. Grey dots correspond to unlabelled data. Green circles and orange triangles are labelled observations from the two classes of observations. The colours of the background show the regions where the algorithm predicts a "green circle" or "orange triangle". Recently labelled instances are highlighted with a bold red border.

interest. We use active learning as a tool to guide the design of an optimal subsampling method, e.g., to estimate a finite population characteristic. We will return to unbiased active learning methods in Chapter 4.2 and 4.3.

3 Optimal subsampling designs

In this Chapter we derive optimal sampling schemes for a general class of estimators, sampling designs, and optimality criteria. Some common notation and general assumptions are introduced in Chapter 3.1. We then start in Chapter 3.2 with linear optimality criteria. More general optimality criteria are considered in Chapter 3.3, and a novel class of expected-distance-minimising optimality criteria introduced in Chapter 3.4. A practical approach to optimal subsampling design by minimising the anticipated covariance matrix under an assisting auxiliary model is discussed in Chapter 3.5. For details on the methods presented in this chapter, we refer to Paper IV. Additional examples, illustrations and applications may be found in Paper I–III.

3.1 Notation and assumptions

First we introduce some notation to put the methods and estimators presented in Chapter 2.1 and 2.2 in a common framework. We consider a p-dimensional finite population characteristic $\boldsymbol{\eta}_0$ for which inference is to be made. For instance, $\boldsymbol{\eta}_0$ may be a vector of finite population totals or function of totals, as in Chapter 2.1, or the solution to an estimating equation, as in Chapter 2.2. Also consider an estimator $\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}$ with the following properties:

$$\mathrm{E}[\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}] = \boldsymbol{\eta}_0 + o(n^{-1/2}), \tag{3.1}$$

$$\mathbf{Cov}(\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}} - \boldsymbol{\eta}_0) = \boldsymbol{\Gamma}(\boldsymbol{\mu}) + o(n^{-1}), \tag{3.2}$$

$$\boldsymbol{\Gamma}(\boldsymbol{\mu}) = \mathbf{A} + \sum_{i \in \mathcal{D}} \frac{1}{\mu_i} \mathbf{B} \boldsymbol{u}_i \boldsymbol{u}_i^{\mathsf{T}} \mathbf{B}^{\mathsf{T}}, \tag{3.3}$$

where $n = \sum_{i \in \mathcal{D}} \mu_i$. As always, $o(\cdot)$ is interpreted elementwise. Under suitable conditions this covers the following classes of estimators introduced in Chapter 2.1 and 2.2:

i) the estimator $\hat{\boldsymbol{t}}_{\boldsymbol{y}}$ for a vector of finite population totals $\boldsymbol{\eta}_0 = \boldsymbol{t}_{\boldsymbol{y}}$, with $\boldsymbol{u}_i = \boldsymbol{y}_i$ and $\mathbf{B} = \mathbf{I}_{p\times p}$,

ii) the estimator $\hat{\boldsymbol{\tau}} = \boldsymbol{h}(\hat{\boldsymbol{t}}_{\boldsymbol{y}})$ for a function of totals $\boldsymbol{\eta}_0 = \boldsymbol{h}(\boldsymbol{t}_{\boldsymbol{y}})$, with $\boldsymbol{u}_i = \boldsymbol{y}_i$ and $\mathbf{B} = \mathbf{J}_{\boldsymbol{h}}(\boldsymbol{t}_{\boldsymbol{y}})$, and

iii) the estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ for a parameter vector $\boldsymbol{\eta}_0 = \boldsymbol{\theta}_0$ defined as the solution to an estimating equation (2.6), with $\boldsymbol{u}_i = \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)$ and $\mathbf{B} = \mathbf{H}(\boldsymbol{\theta}_0)^{-1}$, assuming that $\mathbf{H}(\boldsymbol{\theta}_0)$ is of full rank.

The matrix $\mathbf{A}$ in (3.3) depends on the characteristic of interest and choice of sampling design. For instance, $\mathbf{A}$ is the $p \times p$ zero matrix when sampling is conducted according to a PO-WR design. If, on the other hand, sampling is conducted according to a PO-WOR design, then $\mathbf{A} = -\sum_{i\in\mathcal{D}} \mathbf{B}\boldsymbol{u}_i\boldsymbol{u}_i^{\mathsf{T}}\mathbf{B}^{\mathsf{T}}$ (cf. (2.2) and (2.3)). We may also incorporate the first-stage variance $\mathbf{Cov}(\boldsymbol{\eta}_0 - \boldsymbol{\eta}^*)$ in the matrix $\mathbf{A}$ if super-population inference for an underlying parameter $\boldsymbol{\eta}^*$ is intended (see Chapter 2.2.3).

We note that the approximate covariance matrix $\boldsymbol{\Gamma}(\boldsymbol{\mu})$ of the estimator $\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}$, as given in (3.3), generally depends on unknown full-data characteristics. We will proceed in Chapter 3.2–3.4 as if $\boldsymbol{\Gamma}(\boldsymbol{\mu})$ were known, keeping in mind that the resulting theoretically optimal designs can generally not be found in practice. We refer to Chapter 4 for a discussion on the implementation of optimal subsampling methods in practice.

3.2 Linear optimality criteria

Consider an estimator $\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}$ with approximate covariance matrix $\boldsymbol{\Gamma}(\boldsymbol{\mu})$ given by (3.3), and a family of sampling designs (e.g., PO-WR, PO-WOR or MULTI) of expected size n. We say that a sampling scheme $\boldsymbol{\mu}^*$ is L-optimal for $\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}$ with respect to a $p \times m$ matrix $\mathbf{L}$ if

$$\boldsymbol{\mu}^* = \arg\min_{\boldsymbol{\mu}} \operatorname{tr}(\boldsymbol{\Gamma}(\boldsymbol{\mu})\mathbf{L}\mathbf{L}^{\mathsf{T}}).$$

The minimisation is over the domain of feasible values for the sampling scheme $\boldsymbol{\mu}$. Some special cases include the A-, c- and V-optimality criteria introduced in Chapter 2.3. Other examples of linear optimality criteria will be discussed in Chapter 3.4.

By (3.3) and the cyclic property of the trace, minimising the objective function of the L-optimality criterion is equivalent to the convex constrained optimisation

problem

$$\min_{\boldsymbol{\mu}} \sum_{i\in\mathcal{D}} \mu_i^{-1} c_i, \quad c_i = ||\mathbf{L}^\mathsf{T}\mathbf{B}\boldsymbol{u}_i||_2^2, \tag{3.4}$$

$$\text{subject to } \sum_{i\in\mathcal{D}} \mu_i = n \text{ and } \begin{matrix} \mu_i > 0 & \text{for PO-WR and MULTI,} \\ \mu_i \in (0,1] & \text{for PO-WOR.} \end{matrix} \tag{3.5}$$

The optimal solution for a PO-WR, PO-WOR or MULTI design of (expected) size n is given in Algorithm 3.1. The result is obtained by the Lagrange multiplier method for PO-WR and MULTI designs, and by the Karush-Kuhn-Tucker conditions for PO-WOR (Boyd and Vandenberghe, 2004). Global optimality follows by convexity. See Figure 3.1 for an illustration.

Algorithm 3.1. L-optimal sampling schemes.

INPUT: Index set $\mathcal{D}$, (expected) sample size n, non-zero $p \times m$ matrix $\mathbf{L}$, matrix $\mathbf{B}$ and vectors $\{\boldsymbol{u}_i\}_{i\in\mathcal{D}}$ defined according to (3.3), and family of sampling designs (PO-WR, PO-WOR or MULTI).

Let $c_i = ||\mathbf{L}^\mathsf{T}\mathbf{B}\boldsymbol{u}_i||_2^2$ for all $i \in \mathcal{D}$.
if any $c_i = 0$ **then**
  STOP. Feasible solution does not exist.
else
  Let $\mu_i^* = n\frac{\sqrt{c_i}}{\sum_{j\in\mathcal{D}}\sqrt{c_j}}$ for all $i \in \mathcal{D}$.
  if PO-WOR **then**
    while any $\mu_i^* > 1$ **do**
      Let $\mathcal{E} = \{i \in \mathcal{D} : \mu_i^* \geq 1\}$ and $n_\mathcal{E} = |\mathcal{E}|$.
      Let $\mu_i^* = \begin{cases} 1 & \text{if } i \in \mathcal{E}, \\ (n - n_\mathcal{E})\frac{\sqrt{c_i}}{\sum_{j\in\mathcal{D}\setminus\mathcal{E}}\sqrt{c_j}} & \text{if } i \in \mathcal{D}\setminus\mathcal{E}. \end{cases}$
    end while
  end if
  RETURN optimal sampling scheme $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_N^*)$.
end if

3.3 Non-linear optimality criteria

Now consider a general optimality criterion with corresponding objective function $\Phi : \boldsymbol{S}_+^{p\times p} \to \mathbb{R}$ on the set of real, symmetric, positive semi-definite $p \times p$ matrices. By an optimal sampling scheme $\boldsymbol{\mu}^*$, we mean the following:

Definition 3.1 (Φ-optimality)**.** *Consider a function $\Phi : \boldsymbol{S}_+^{p\times p} \to \mathbb{R}$ that is monotone for Loewner's ordering, i.e., such that* (2.8) *holds. Also, consider a family of unequal*

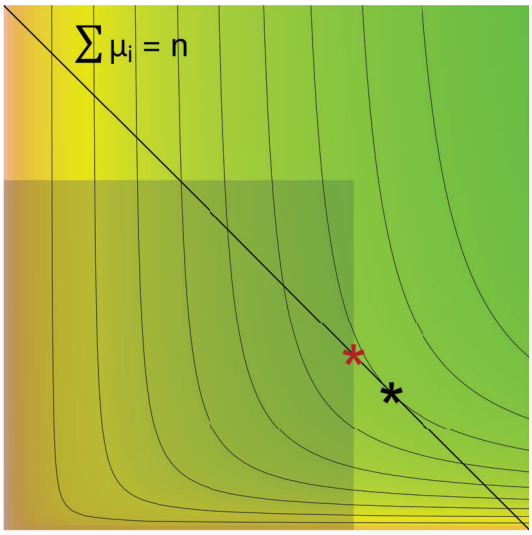


Figure 3.1. Illustration of the convex constrained optimisation problem (3.4)–(3.5). For a PO-WR or MULTI design, the constrained minimum (black star) lies where the $(N-1)$-dimensional hyperplane $\sum_{i\in\mathcal{D}} \mu_i = n$ tangents the level set of the function $\sum_i \mu_i^{-1} c_i$. If, for PO-WOR, this solution does not lie within the hypercube $(0,1]^N$ (shaded region), the sampling probabilities are set to $\min\{\mu_i, 1\}$ and the regained probability mass reallocated among the instances with $\mu_i < 1$. The procedure is repeated until a feasible solution is obtained (red star) (Algorithm 3.1).

probability sampling designs (e.g, PO-WR, PO-WOR, or MULTI) of (expected) size n, indexed by the sampling scheme $\boldsymbol{\mu}$, and let $\mathcal{M}_n$ denote the corresponding domain of $\boldsymbol{\mu}$. We say that a sampling scheme $\boldsymbol{\mu}^*$ is Φ-optimal for $\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}$ if

$$\boldsymbol{\mu}^* = \underset{\boldsymbol{\mu}\in\mathcal{M}_n}{\arg\min}\, \Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu})), \tag{3.6}$$

where $\boldsymbol{\Gamma}(\boldsymbol{\mu})$ is the approximate covariance matrix of $\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}$, as given in (3.3).

Finding a Φ-optimal sampling scheme reduces to a non-linear, possibly non-convex, restricted optimisation problem over an $(N-1)$-dimensional hyperplane in $\mathbb{R}^N$. A key observation to obtain Φ-optimality is to note the following: if $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu}))$ is twice differentiable at some point $\boldsymbol{\mu}^*$, and the Hessian is positive-semi-definite, then the objective function behaves locally like the L-optimality criterion in a neighbourhood of $\boldsymbol{\mu}^*$. Indeed, we have for the PO-WR, PO-WOR and MULTI designs that

$$\frac{\partial\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu}))}{\partial\mu_i} = -\mu_i^{-2}\left|\left|\mathbf{L}(\boldsymbol{\mu})^\mathsf{T}\mathbf{B}\boldsymbol{u}_i\right|\right|_2^2,$$

provided that the derivative exists. Here $\mathbf{L}(\boldsymbol{\mu})$ is a real matrix such that $\mathbf{L}(\boldsymbol{\mu})\mathbf{L}(\boldsymbol{\mu})^\mathsf{T} = \phi(\boldsymbol{\Gamma}(\boldsymbol{\mu}))$, and $\phi(\mathbf{U}) = \frac{\partial\Phi(\mathbf{U})}{\partial\mathbf{U}}$ is the $p\times p$ matrix derivative of

Φ with respect to its matrix argument. Examples of objective functions and their corresponding matrix derivatives for some common optimality criteria include:

i) The D-optimality objective function $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu})) = \log\det(\boldsymbol{\Gamma}(\boldsymbol{\mu}))$, which is differentiable with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu})) = \boldsymbol{\Gamma}(\boldsymbol{\mu})^{-1}$, provided that $\boldsymbol{\Gamma}(\boldsymbol{\mu})$ is of full rank.

ii) The E-optimality objective function $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu})) = \lambda_{\max}(\boldsymbol{\Gamma}(\boldsymbol{\mu}))$, which is differentiable with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu})) = \boldsymbol{v}_{\boldsymbol{\mu}}\boldsymbol{v}_{\boldsymbol{\mu}}^{\mathsf{T}}$, provided that $\lambda_{\max}(\boldsymbol{\Gamma}(\boldsymbol{\mu}))$ has multiplicity 1. Here $\boldsymbol{v}_{\boldsymbol{\mu}}$ is the eigenvector pertaining to the maximal eigenvalue of $\boldsymbol{\Gamma}(\boldsymbol{\mu})$

iii) The L-optimality objective function $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu})) = \mathrm{tr}(\boldsymbol{\Gamma}(\boldsymbol{\mu})\mathbf{L}\mathbf{L}^{\mathsf{T}})$, which is differentiable with respect to $\boldsymbol{\mu}$, and $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu})) = \mathbf{L}\mathbf{L}^{\mathsf{T}}$. In particular, this holds for A-optimality with $\mathbf{L} = \mathbf{I}_{p\times p}$ and c-optimality with $\mathbf{L} = \mathbf{c}$.

See Wand (2002), Petersen and Pedersen (2012) and Magnus and Neudecker (2019) for details on matrix calculus and matrix differentiation. Optimality conditions for a general Φ-optimality criterion may now be derived by the Lagrange multiplier method and Karush-Kuhn-Tucker conditions, in analogy with the results for L-optimality (Proposition 3.1).

Proposition 3.1 (Φ-optimality conditions)**.** *Consider the family of PO-WR, PO-WOR or MULTI designs of (expected) size n. Also consider a function $\Phi : \boldsymbol{S}_{+}^{p\times p} \to \mathbb{R}$ such that Φ is monotone for Loewner's ordering. Let $\boldsymbol{\Gamma}(\boldsymbol{\mu})$, $\mathbf{B}$ and $\{\boldsymbol{u}_i\}_{i\in\mathcal{D}}$ be defined according to* (3.3)*, and assume that $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu}))$ is differentiable with respect to $\boldsymbol{\mu}$ in a neighbourhood of some point $\boldsymbol{\mu}^*$. Let $\boldsymbol{\phi}(\mathbf{U}) = \frac{\partial\Phi(\mathbf{U})}{\partial\mathbf{U}}$, and $\mathbf{L}(\boldsymbol{\mu}^*)$ be a real matrix such that $\mathbf{L}(\boldsymbol{\mu}^*)\mathbf{L}(\boldsymbol{\mu}^*)^{\mathsf{T}} = \boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu}^*))$. Finally, let*

$$c_i = \left|\left|\mathbf{L}(\boldsymbol{\mu}^*)^{\mathsf{T}}\mathbf{B}\boldsymbol{u}_i\right|\right|_2^2.$$

Then the following holds:

a) $\boldsymbol{\mu}^*$ is a stationary point of $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu}))$ for a PO-WR or MULTI design of size n if

$$\mu_i^* = n\frac{\sqrt{c_i}}{\sum_{j\in\mathcal{D}}\sqrt{c_j}} \quad \textit{for all } i \in \mathcal{D}.$$

b) $\boldsymbol{\mu}^*$ *is a stationary point of* $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu}))$ *for a PO-WOR design of size* n *if*

$$\begin{aligned} \mu_i^* &\leq 1 && \text{for all } i \in \mathcal{D}, \\ \mu_i^* &= (n - n_{\mathcal{E}}) \frac{\sqrt{c_i}}{\sum_{j \in \mathcal{D} \setminus \mathcal{E}} \sqrt{c_j}} && \text{for all } i \in \mathcal{D} \setminus \mathcal{E}, \\ \sqrt{c_i} &\geq \sqrt{c_j}/\mu_j^* && \text{for all } i \in \mathcal{E} \text{ and } j \in \mathcal{D} \setminus \mathcal{E}, \end{aligned}$$

where $\mathcal{E} = \{i \in \mathcal{D} : \mu_i^* = 1\}$ *and* $n_{\mathcal{E}} = |\mathcal{E}|$.

Consequently, if $\boldsymbol{\mu}^*$ *satisfies the optimality conditions according to a) or b), and* $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu}))$ *is convex in* $\boldsymbol{\mu}$*, then* $\boldsymbol{\mu}^*$ *is the global minimiser of* $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu}))$.

Based on the above results and observations, in Algorithm 3.2 we present a fixed-point iteration method to find optimal sampling schemes for non-linear optimality criteria. The algorithm takes an initial sampling scheme as input, and solves a series of convex optimisation problems by a local approximation of the objective function as linear optimality criterion. The algorithm is terminated for convergence when the relative improvement of the objective function between two consecutive iterations is less than some pre-specified tolerance level ϵ (e.g., $\epsilon = 10^{-3}$). The algorithm may also be terminated for divergence if the value of the objective function increases between the iterations. For L-optimality, the method is exact and terminates within a single iteration. Beyond L-optimality, the algorithm need not converge, and even if it does, it need not converge to a global optimum unless the problem is convex. Both the D- and L-optimality criteria are convex in $\boldsymbol{\mu}$, which implies that global optimality can be deduced.

3.4 Expected-distance-minimising designs

Recall the overall aim of data subsampling as introduced in Chapter 1; to find an approximate solution to an original intractable problem of estimating a finite population characteristic or full-data parameter $\boldsymbol{\eta}_0$. A natural target of optimal design in this context is to minimise the expected distance $\mathrm{E}[d(\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}})]$ of the estimator $\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}$ from the full-data parameter $\boldsymbol{\eta}_0$, for some suitable statistical distance function $d : \boldsymbol{\Omega} \to \mathbb{R}_+$ on the set $\boldsymbol{\Omega}$ of possible values for the estimator $\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}$.

We require a statistical distance function $d(\boldsymbol{\eta})$ that is greater than or equal to zero for all $\boldsymbol{\eta}$, with equality only for $\boldsymbol{\eta} = \boldsymbol{\eta}_0$. For analytical and computational tractability we also require the distance function to be twice differentiable, and

Algorithm 3.2. Fixed point iteration.

INPUT: Index set $\mathcal{D}$, (expected) sample size n, optimality criterion Φ, matrix $\mathbf{B}$ and vectors $\{\boldsymbol{u}_i\}_{i\in\mathcal{D}}$ defined according to (3.3), initial sampling scheme $\boldsymbol{\mu}^{(0)}$, family of sampling designs (PO-WR, PO-WOR or MULTI), maximal number of iterations T, tolerance parameter $\epsilon > 0$.

for t = 1, . . . , T **do**
  Let $\mathbf{L}_t$ be a matrix such that $\mathbf{L}_t\mathbf{L}_t^\intercal = \boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu}^{(t-1)}))$.
  Let $c_i = ||\mathbf{L}_t^\intercal \mathbf{B}\boldsymbol{u}_i||_2^2$ for all $i \in \mathcal{D}$.
  if any $c_i = 0$ **then**
    STOP. Unfeasible solution encountered during iteration.
  else
    Find L-optimal sampling scheme $\boldsymbol{\mu}^{(t)}$ with respect to $\mathbf{L} = \mathbf{L}_t$ according to Algorithm 3.1.
    if value of objective function increased **then**
      STOP. Algorithm diverged.
    else if relative improvement of the objective function $< \epsilon$ **then**
      Algorithm converged. RETURN $\boldsymbol{\mu}^* = \boldsymbol{\mu}^{(t)}$.
    end if
  end if
end for

denote by $\mathbf{H}_d(\boldsymbol{\eta}) = \frac{\partial^2 d(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\intercal}$ the Hessian matrix of $d(\boldsymbol{\eta})$. By a second order Taylor expansion around $\boldsymbol{\eta}_0$, we have that

$$\begin{aligned} d(\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}) = d(\boldsymbol{\eta}_0) + \nabla d(\boldsymbol{\eta})^\intercal\big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0}(\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}} - \boldsymbol{\eta}_0) \\ + \frac{1}{2}(\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}} - \boldsymbol{\eta}_0)^\intercal \mathbf{H}_d(\boldsymbol{\eta}_0)(\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}} - \boldsymbol{\eta}_0) + o_p(||(\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}} - \boldsymbol{\eta}_0)||_2^2), \end{aligned}$$

where, by definition of $d(\boldsymbol{\eta})$, the first two terms are zero. Under mild assumptions, including (3.1) and (3.2), it follows that

$$\mathrm{E}[d(\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}})] = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Gamma}(\boldsymbol{\mu})\mathbf{H}_d(\boldsymbol{\eta}_0)\right) + o(n^{-1}).$$

Hence, we define a class of expected-distance-minimising optimality criteria as follows:

Definition 3.2 (d-optimality)**.** *Consider a function $d : \boldsymbol{\Omega} \to \mathbb{R}_+$ such that $d(\boldsymbol{\eta}) = 0$ if and only if $\boldsymbol{\eta} = \boldsymbol{\eta}_0$. Assume that $d(\boldsymbol{\eta})$ is twice differentiable in a neighbourhood of $\boldsymbol{\eta}_0$, and that $\mathbf{H}_d(\boldsymbol{\eta}_0)$ is non-zero. Also, consider a family of unequal probability sampling designs (e.g, PO-WR, PO-WOR, or MULTI) of (expected) size n, indexed by the sampling scheme $\boldsymbol{\mu}$, and denote by $\mathcal{M}_n$ the corresponding domain of $\boldsymbol{\mu}$. We say*

that a sampling scheme $\boldsymbol{\mu}^*$ is d-optimal if

$$\boldsymbol{\mu}^* = \underset{\boldsymbol{\mu} \in \mathcal{M}_n}{\arg\min}\, \mathrm{tr}\left(\boldsymbol{\Gamma}(\boldsymbol{\mu}) \mathbf{H}_d(\boldsymbol{\eta}_0)\right).$$

We denote this optimality criterion as d-optimality for *distance*, which should not be confused with the D-optimality criterion introduced in Chapter 2.3. We note that d-optimality is equivalent to L-optimality with respect to a matrix $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}^\mathsf{T} = \mathbf{H}_d(\boldsymbol{\eta}_0)$. Indeed, any (differentiable) Φ-optimality criterion may be viewed as a d-optimality criterion, and vice versa. For instance, A-optimality is equivalent to d-optimality when $d(\boldsymbol{\eta}) = ||\boldsymbol{\eta} - \boldsymbol{\eta}_0||_2^2$, the squared Euclidean distance. Beyond linear optimality criteria, the induced distance function may be implicit and depend on the Φ-optimal sampling scheme $\boldsymbol{\mu}^*$. For instance, E-optimality is equivalent to d-optimality when $d(\boldsymbol{\eta}) = ||\boldsymbol{v}_{\boldsymbol{\mu}^*}^\mathsf{T}(\boldsymbol{\eta} - \boldsymbol{\eta}_0)||_2^2$, where $\boldsymbol{v}_{\boldsymbol{\mu}^*}$ is an eigenvector pertaining to the largest eigenvalue of $\boldsymbol{\Gamma}(\boldsymbol{\mu}^*)$ and $\boldsymbol{\mu}^*$ the corresponding E-optimal sampling scheme. In this case the distance function for the d-optimality criterion can only be evaluated if the E-optimal sampling scheme is known.

To illustrate the use of the d-optimality criterion, consider the following statistical distance functions that arise naturally in data subsampling applications and are commonly encountered in statistics:

i) **Kullback-Leibler divergence**: Consider a random vector $\boldsymbol{Y}$ with probability density function $f_{\boldsymbol{\eta}}(\boldsymbol{y})$ and cumulative distribution function $F_{\boldsymbol{\eta}}(\boldsymbol{y})$. Let $\mathcal{Y}$ denote the domain of $\boldsymbol{Y}$. The Kullback-Leibler divergence of $f_{\boldsymbol{\eta}}$ from $f_{\boldsymbol{\eta}_0}$ is defined as $\mathrm{KL}\big(f_{\boldsymbol{\eta}_0} || f_{\boldsymbol{\eta}}\big) = \int_{\mathcal{Y}} \log \frac{f_{\boldsymbol{\eta}_0}(\boldsymbol{y})}{f_{\boldsymbol{\eta}}(\boldsymbol{y})} dF_{\boldsymbol{\eta}_0}(\boldsymbol{y})$. To allow for covariates, we define the Kullback-Leibler distance of $\boldsymbol{\eta}$ from $\boldsymbol{\eta}_0$ as $d_{\mathrm{KL}}(\boldsymbol{\eta}) = \sum_{i \in \mathcal{D}} \int_{\mathcal{Y}} \log \frac{f_{\boldsymbol{\eta}_0}(\boldsymbol{y}|\boldsymbol{x}_i)}{f_{\boldsymbol{\eta}}(\boldsymbol{y}|\boldsymbol{x}_i)} dF_{\boldsymbol{\eta}_0}(\boldsymbol{y}|\boldsymbol{x}_i)$.

ii) **Empirical risk distance**: Assume that $\boldsymbol{\eta}_0$ is defined as the minimiser of an empirical risk $\ell_0(\boldsymbol{\eta})$ as in (2.4). A natural measure for the distance of a parameter value $\boldsymbol{\eta}$ from $\boldsymbol{\eta}_0$ is through the difference in the attained value of the empirical risk. We define the empirical risk distance of $\boldsymbol{\eta}$ from $\boldsymbol{\eta}_0$ as $d_{\mathrm{ER}}(\boldsymbol{\eta}) = \ell_0(\boldsymbol{\eta}) - \ell_0(\boldsymbol{\eta}_0)$.

iii) **Mahalanobis distance**: Consider a probability distribution on $\mathbb{R}^p$ with mean vector $\boldsymbol{\gamma}$ and covariance matrix $\boldsymbol{\Sigma}$. The Mahalanobis distance of a point $\boldsymbol{\eta} \in \mathbb{R}^p$ from the mean $\boldsymbol{\gamma}$ is then given by $\sqrt{(\boldsymbol{\eta} - \boldsymbol{\gamma})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\eta} - \boldsymbol{\gamma})}$. We define the squared Mahalanobis distance of $\boldsymbol{\eta}$ from $\boldsymbol{\eta}_0$ with respect to a real, symmetric, positive definite dispersion matrix $\boldsymbol{\Sigma}$ as $d_{\boldsymbol{\Sigma}}(\boldsymbol{\eta}) = (\boldsymbol{\eta} - \boldsymbol{\eta}_0)^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\eta} - \boldsymbol{\eta}_0)$.

An illustration is provided in Figure 3.2. We define d_{ER}-, d_{KL}- and $d_{\boldsymbol{\Sigma}}$-optimality accordingly, i.e., as d-optimality with the distance function taken as indicated by the subscript. Four natural choices of the dispersion matrix $\boldsymbol{\Sigma}$ for the Mahalanobis distance are:

iii.a) $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}(\boldsymbol{\mu})$, the approximate covariance matrix of $\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}$,

iii.b) $\boldsymbol{\Sigma} = \boldsymbol{\mathcal{I}}(\boldsymbol{\eta}_0)^{-1}$, where $\boldsymbol{\mathcal{I}}(\boldsymbol{\eta}_0)$ is defined for parametric model as the expected value of the Hessian of the full-data empirical risk, i.e., the expected Fisher information matrix,

iii.c) $\boldsymbol{\Sigma} = \mathbf{H}(\boldsymbol{\eta}_0)^{-1}$, where $\mathbf{H}(\boldsymbol{\eta}_0)$ is defined for an empirical risk minimiser $\boldsymbol{\eta}_0$ as the Hessian of the full-data empirical risk, i.e., the observed Fisher information matrix, and

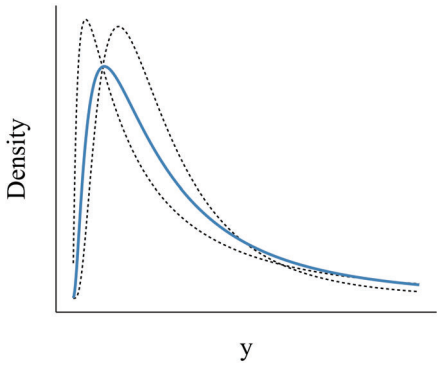
iii.d) $\boldsymbol{\Sigma} = \mathbf{B}(\mathbf{U}^{\mathsf{T}}\mathbf{U})\mathbf{B}$, where $\mathbf{U}$ is the matrix with rows $\boldsymbol{u}_i^{\mathsf{T}}$, $i = 1, \ldots, N$, and $\mathbf{B}, \boldsymbol{u}_i$ are defined according to (3.3).

We note that $d_{\boldsymbol{\Sigma}}$-optimality with $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}(\boldsymbol{\mu})$ corresponds to D-optimality, $\boldsymbol{\Sigma} = \boldsymbol{\mathcal{I}}(\boldsymbol{\eta}_0)^{-1}$ to d_{KL}-optimality, and $\boldsymbol{\Sigma} = \mathbf{H}(\boldsymbol{\eta}_0)^{-1}$ to d_{ER}-optimality. The last choice of dispersion matrix, $\boldsymbol{\Sigma} = \mathbf{B}(\mathbf{U}^{\mathsf{T}}\mathbf{U})\mathbf{B}$, arises from M-estimation theory as the empirical covariance matrix of the estimator $\boldsymbol{\eta}_0$, seen as an estimator of an underlying super-population parameter $\boldsymbol{\eta}^*$ (see, e.g., Stefanski and Boos, 2002). Apart from $d_{\boldsymbol{\Sigma}}$-optimality with $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}(\boldsymbol{\mu})$, all above-mentioned optimality criteria are examples of linear optimality criteria. Hence, an optimal sampling scheme may be found using the methods described in Chapter 3.2.

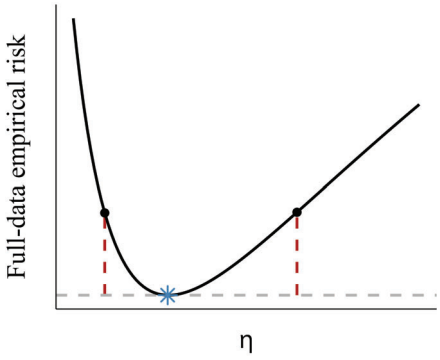
In addition to their appealing geometric and statistical interpretation, the expected-distance-minimising optimality criteria introduced above have two desirable properties: computational tractability and parameterisation invariance. Indeed, belonging to the class of linear optimality criteria, the d_{ER}-, d_{KL}- and $d_{\boldsymbol{\Sigma}}$-optimality criteria have simple solutions for the optimal sampling schemes according to Algorithm 3.1. Moreover, these optimality criteria are invariant under a change of parameterisation $\boldsymbol{\eta} \mapsto \boldsymbol{g}(\boldsymbol{\eta})$, for a one-to-one differentiable transformation $\boldsymbol{g} : \boldsymbol{\Omega} \to \boldsymbol{\Omega}$. With a few exceptions and pathological examples, the same also holds for non-singular affine transformations of the data. See Paper IV for further details.

3.5 Auxiliary-variable-assisted designs

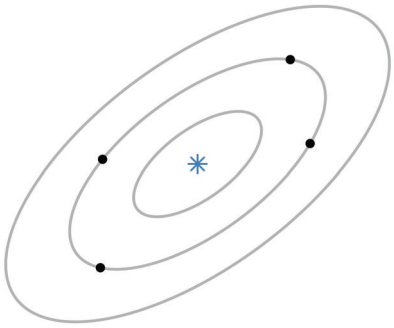
Thus far, we have assumed the approximate covariance matrix (3.3) to be known. In practice, this matrix depends on unknown full-data characteristics.



A: Kullback-Leibler distance.
The solid blue line shows the density of a parametric model $f_{\boldsymbol{\eta}}(y)$ evaluated at the full-data parameter $\boldsymbol{\eta}_0$. The dashed black lines show the densities for two other values of the parameter $\boldsymbol{\eta}$. The Kullback-Leibler distance from $\boldsymbol{\eta}_0$ to $\boldsymbol{\eta}$ is given by $d_{\mathrm{KL}}(\boldsymbol{\eta}) = \int_{\mathcal{Y}} \log \frac{f_{\boldsymbol{\eta}_0}(y)}{f_{\boldsymbol{\eta}}(y)} f_{\boldsymbol{\eta}_0}(\boldsymbol{y}) dy$.



B: Empirical risk distance.
The solid black line shows the full-data empirical risk $\ell_0(\boldsymbol{\eta})$ as a function of the parameter $\boldsymbol{\eta}$. The empirical risk distance from the minimiser $\boldsymbol{\eta}_0$ to $\boldsymbol{\eta}$ is defined as the difference in the attained full-data empirical risk, i.e., $d_{\mathrm{ER}}(\boldsymbol{\eta}) = \ell_0(\boldsymbol{\eta}) - \ell_0(\boldsymbol{\eta}_0)$. This is illustrated in the figure by the dashed red lines.



C: Mahalanobis distance.
The grey lines show the level curves of the Mahalanobis distance function $d_{\boldsymbol{\Sigma}}(\boldsymbol{\eta}) = \sqrt{(\boldsymbol{\eta} - \boldsymbol{\eta}_0)^{\intercal} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\eta} - \boldsymbol{\eta}_0)}$ from the full-data parameter $\boldsymbol{\eta}_0$ to $\boldsymbol{\eta}$ with respect to a positive definite dispersion matrix $\boldsymbol{\Sigma}$.

Figure 3.2. Visualisation of the Kullback-Leibler distance (**A**), empirical risk distance (**B**), and Mahalanobis distance functions (**C**). In each panel, the black dots/dashed curves (parameter values $\boldsymbol{\eta}$) are equally distant from the blue star/blue solid curve (full-data parameter $\boldsymbol{\eta}_0$) in Kullback-Leibler-, empirical risk-, or Mahalanobis-sense, respectively.

For instance, when studying a finite population characteristic $\boldsymbol{\tau} = \boldsymbol{h}(\boldsymbol{t}_{\boldsymbol{y}})$, the approximate covariance matrix $\boldsymbol{\Gamma}(\boldsymbol{\mu})$ depends on the study variables $\boldsymbol{y}_i$. For a non-linear function $\boldsymbol{h}(\boldsymbol{t}_{\boldsymbol{y}})$ it further depends on the value of the total $\boldsymbol{t}_{\boldsymbol{y}}$. For inference regarding the full-data empirical risk minimiser (2.4), the approximate covariance matrix (3.3) may depend on the covariates $\boldsymbol{x}_i$, responses $\boldsymbol{y}_i$, and full-data parameter $\boldsymbol{\theta}_0$. However, if such information were available at the design stage, subsampling would not be needed in the first place. In this section we show how optimal subsampling schemes can be found in practice by utilisation of auxiliary information about the members of $\mathcal{D}$ that is available prior to subsampling.

In addition to the data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i\in\mathcal{D}}$, we assume the existence of a collection of auxiliary variables $\{\boldsymbol{z}_i\}_{i\in\mathcal{D}}$, which are *a priori* available for all members $i \in \mathcal{D}$ (Figure 1.1, Chapter 1). Depending on context, the auxiliary variables may include some of the variables in $\boldsymbol{x}_i$ and/or some of the variables in $\boldsymbol{y}_i$. In Paper I, for instance, we consider an active learning problem were the predictors $\boldsymbol{x}_i$ are assumed to be known for all instances in a large pool of potential training examples, whereas the outcomes y_i are affordable to observe only for a subset. Hence, we have $\boldsymbol{x}_i = \boldsymbol{z}_i$ in this case. In Paper II we consider a naturalistic driving study to analyse risk factors for a safety critical event in a case-control setting. In this setting the outcome is known and hence included in the auxiliary variables. In addition, the auxiliary variables include proxies for the explanatory variables, derived from automatic recordings of the vehicle kinematics. In Paper III we consider a large computer experiment where the outcomes are computationally expensive to observe. In this case the auxiliary variables are the inputs to the experiment, which are under immediate control of the investigator. By utilising such auxiliary information, we may obtain an approximate solution to the optimisation problem (3.6), under an assisting auxiliary model for the unknown values of the study variables of interest.

We introduce a collection of random variables $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i\in\mathcal{D}}$ to describe our uncertainty in the unknown values of the data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i\in\mathcal{D}}$. For any variable also included in $\boldsymbol{z}_i$, we may associate a degenerate (deterministic) distribution with the corresponding component of $(\mathbf{X}_i, \mathbf{Y}_i)$ conditioned on $\boldsymbol{z}_i$. We also assume that we have a preliminary estimate $\tilde{\boldsymbol{\eta}}_0$ of the full-data parameter $\boldsymbol{\eta}_0$, and an auxiliary model $f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z})$ for the conditional distribution of the random variables $(\mathbf{X}_i, \mathbf{Y}_i)$ given auxiliary variables $\boldsymbol{z}_i$. Such information may be available from domain knowledge, previous studies, a pilot sample, or a combination of those. In Chapter 4 we will discuss how such information can be acquired gradually during the subsampling process. Following Isaki and Fuller (1982), we define the anticipated covariance matrix of an estimator $\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}$ as follows:

Definition 3.3 (Anticipated covariance)**.** *Consider a data triplet* $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{z}_i)\}_{i\in\mathcal{D}}$,

where $(\mathbf{X}_i, \mathbf{Y}_i)$ *is a random vector and* $\boldsymbol{z}_i$ *are known for all* $i \in \mathcal{D}$*. Also consider a preliminary estimate* $\tilde{\boldsymbol{\eta}}_0$ *of the full-data parameter* $\boldsymbol{\eta}_0$*, and a model* $f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z})$ *for the conditional distribution of* $(\mathbf{X}_i, \mathbf{Y}_i)$ *given auxiliary variables* $\boldsymbol{z}_i$*. The anticipated covariance matrix of* $\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}$ *is defined as*

$$\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\mu}; \tilde{\boldsymbol{\eta}}_0) = \mathrm{E}_{(\boldsymbol{x},\boldsymbol{y})\sim f(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})}[\boldsymbol{\Gamma}(\boldsymbol{\mu})]\big|_{\boldsymbol{\eta}_0=\tilde{\boldsymbol{\eta}}_0}.$$

All results in Chapter 3.2–3.4 may now be restated for Φ-optimality with respect to the anticipated covariance matrix $\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\mu}; \tilde{\boldsymbol{\eta}}_0)$ instead of the approximate covariance matrix $\boldsymbol{\Gamma}(\boldsymbol{\mu})$. Under weak assumptions on the model $f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z})$ that allow us to replace the order of integration and differentiation, all that changes is that the coefficients c_i in Algorithm 3.2 are replaced by their corresponding expectations

$$\tilde{c}_i := \mathrm{E}_{(\boldsymbol{x},\boldsymbol{y})\sim f(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})}\left[C_i|\boldsymbol{z}_i\right], \quad C_i = \left|\left|\mathbf{L}_t^\mathsf{T}\mathbf{B}\boldsymbol{u}_i\right|\right|_2^2, \tag{3.7}$$

which, if $\boldsymbol{\Gamma}(\boldsymbol{\mu})$ depends on the full-data parameter $\boldsymbol{\eta}_0$, is evaluated at a preliminary parameter estimate $\tilde{\boldsymbol{\eta}}_0$. Here C_i is a function of the random variables $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i\in\mathcal{D}}$, and $\mathbf{L}_t$ a matrix such that $\mathbf{L}_t\mathbf{L}_t^\mathsf{T} = \boldsymbol{\phi}(\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\mu}^{(t-1)}; \tilde{\boldsymbol{\eta}}_0))$.

4 Sequential optimal design

To implement the auxiliary-variable-assisted optimal subsampling methods of Chapter 3.5, we need a preliminary estimate of the characteristic of interest and an auxiliary model for the unknowns. In Section 4.1 we show how this can be accomplished by sequential optimal design, were the information needed to evaluate the expectation in (3.7) is acquired gradually during the sampling process. Two implementations of the sequential optimal design method for unbiased active learning and active sampling for finite population inference are discussed in Chapter 4.2 and 4.3, respectively. A martingale central limit theorem, establishing consistency and asymptotic normality of the resulting active sampling estimator, is presented in Chapter 4.4.

4.1 A sequential subsampling algorithm

Assume that subsampling is conducted in K interdependent stages as follows. Let $\{\boldsymbol{S}_k\}_{k=1}^K$ be a sequence of dependent random vectors $\boldsymbol{S}_k = (S_{k1}, \ldots, S_{kN})$ on the N-dimensional non-negative integer lattice. The variable S_{ki} describes the number of times an instance $i \in \mathcal{D}$ is selected by the sampling mechanism in stage k. Denote by $\mu_{ki} = \mathrm{E}[S_{ki}|\boldsymbol{S}_1, \ldots \boldsymbol{S}_{k-1}]$ the corresponding mean number of selections, conditioned on the preceding sampling stages. Since this is now a random variable, we require $\mu_{ki} > 0$ with probability 1 for all k, i. Let $n_k < N$ be the (expected) size of the subsample selected at stage k, so that $\sum_{i\in\mathcal{D}} \mu_{ki} = n_k$, and let $m_k = \sum_{j=1}^k n_j$ be the cumulative sample size after k sampling stages. An estimator for a function of totals $\boldsymbol{\tau} = \boldsymbol{h}(\boldsymbol{t_y})$ after k sampling stages may be defined as

$$\hat{\boldsymbol{\tau}}^{(k)} = \boldsymbol{h}(\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)}) \tag{4.1}$$

with

$$\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)} = m_k^{-1} \sum_{j=1}^{k} n_j \hat{\boldsymbol{t}}_{\boldsymbol{y},j}, \quad \hat{\boldsymbol{t}}_{\boldsymbol{y},j} = \sum_{i \in \mathcal{D}} S_{ji} w_{ji} \boldsymbol{y}_i, \quad w_{ji} = 1/\mu_{ji}.$$

Here $\hat{\boldsymbol{t}}_{\boldsymbol{y},j}$ is an unbiased Hansen-Hurwitz estimator of the total $\boldsymbol{t}_{\boldsymbol{y}}$ from the sample obtained at stage j, and $\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)}$ a pooled estimator calculated from the first k subsamples. Similarly, an estimator for the full-data empirical risk minimiser $\boldsymbol{\theta}_0$, defined by (2.4)–(2.6), may be obtained as

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}^{(k)} = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \boldsymbol{\Omega}} \hat{\ell}_{\boldsymbol{\mu}}^{(k)}(\boldsymbol{\theta}), \tag{4.2}$$

$$\hat{\ell}_{\boldsymbol{\mu}}^{(k)}(\boldsymbol{\theta}) = m_k^{-1} \sum_{j=1}^{k} n_j \hat{\ell}_{\boldsymbol{\mu},j}(\boldsymbol{\theta}), \quad \hat{\ell}_{\boldsymbol{\mu},j}(\boldsymbol{\theta}) = \sum_{i \in \mathcal{D}} S_{ji} w_{ji} \ell_i(\boldsymbol{\theta}), \quad w_{ji} = 1/\mu_{ji}.$$

Now consider a characteristic $\boldsymbol{\eta}_0 = \boldsymbol{h}(\boldsymbol{t}_{\boldsymbol{y}})$ or $\boldsymbol{\eta}_0 = \boldsymbol{\theta}_0$, and let $\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}^{(k)}$ be a corresponding estimator on the form (4.1) or (4.2). Assume, in analogy with (3.1)–(3.3), that

$$\begin{aligned} \mathrm{E}[\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}^{(k)}] &= \boldsymbol{\eta}_0 + o(m_k^{-1/2}), \\ \mathbf{Cov}(\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}^{(k)} - \boldsymbol{\eta}_0) &= \boldsymbol{\Gamma}_k(\boldsymbol{\mu}_{1:k}) + o_p(m_k^{-1}), \\ \boldsymbol{\Gamma}_k(\boldsymbol{\mu}_{1:k}) &= m_k^{-2} \sum_{j=1}^{k} n_j^2 \left(\mathbf{A}_j + \sum_{i \in \mathcal{D}} \mu_{ji}^{-1} \mathbf{B} \boldsymbol{u}_i \boldsymbol{u}_i^{\mathsf{T}} \mathbf{B}^{\mathsf{T}} \right), \end{aligned}$$

where $\boldsymbol{\mu}_{1:k} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)$ and $\boldsymbol{\mu}_j = (\mu_{j1}, \ldots, \mu_{jN}), j = 1, \ldots, k$. In other words, we assume that $\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}^{(K)}$ is design-consistent for $\boldsymbol{\eta}_0$, and that $m_K \boldsymbol{\Gamma}_K(\boldsymbol{\mu}_{1:K})$ converges elementwise in probability to the asymptotic covariance matrix of $\sqrt{m_K}(\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}^{(K)} - \boldsymbol{\eta}_0)$ as the total sample size m_K tends to infinity. A formal justification is provided in Chapter 4.4 and in the appendix of Paper III.

We may now proceed with optimal subsampling in an iterative fashion according to Algorithm 4.1. In each iteration, we select a random sample and retrieve the corresponding data $(\boldsymbol{x}_i, \boldsymbol{y}_i)$. As more data is observed, we update our estimate of the target characteristic $\boldsymbol{\eta}_0$ according to (4.1) or (4.2). We also update our auxiliary model $f(\boldsymbol{x}, \boldsymbol{y} | \boldsymbol{z})$ for the unknown study variables given the auxiliary variables $\boldsymbol{z}_i$. With this information at hand, we may derive a Φ-optimal sampling scheme $\boldsymbol{\mu}_k^*$ for the next sampling stage. This is done by minimising the Φ-optimality criterion with respect to the expectation of $\boldsymbol{\Gamma}_k(\boldsymbol{\mu}_{1:k})$ under the current model $f_{k-1}(\boldsymbol{x}, \boldsymbol{y} | \boldsymbol{z})$ and evaluated under the current estimate $\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}^{(k-1)}$.

Two specific examples of the sequential subsampling method for unbiased

active learning and active sampling for finite population inference are described in Chapter 4.2 and 4.3, respectively.

Algorithm 4.1. Sequential optimal subsampling designs.

INPUT: Index set $\mathcal{D}$, optimality criterion Φ, family of sampling designs (PO-WR, PO-WOR or MULTI), number of sampling stages K, batch sizes $\{n_k\}_{k=1}^{K}$.

1: **for** $k = 1, \ldots, K$ **do**
2: Calculate Φ-optimal sampling scheme.
3: Select a random subsample of size n_k.
4: Retrieve data $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ for the selected instances.
5: Estimate the target parameter $\boldsymbol{\eta}_0$.
6: Update the auxiliary model $f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z})$.
7: Evaluate performance/precision.
8: STOP if sufficient precision in reached. ELSE continue.
9: **end for**

4.2 Unbiased active learning

Consider, as in Chapter 2.4, the problem of training a prediction model $f_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i)$ for some outcome y_i given observed inputs $\boldsymbol{x}_i$. The predictors $\boldsymbol{x}_i$ are assumed to be known for all elements i in a large pool $\mathcal{D}$ of potential training examples, whereas the outcomes y_i are affordable to observe only for a subset of size $n \ll N$. An active learning algorithm utilise information about the known predictors to select which of the outcomes to observe for optimal performance.

A drawback of traditional active learning methods is the sensitivity to model misspecification; see the discussion in Chapter 2.2.3 and 2.4. Unbiased active learning offers a promising solution to this problem (Bach, 2007; Chu et al., 2011; Ganti and Gray, 2012; Farquhar et al., 2021). Using unequal probability sampling and inverse probability weighting, an unbiased estimator of the full-data empirical risk is obtained. Consequently, consistent estimators and predictions may be derived. This holds true even under the realistic assumption of model misspecification (cf. Skinner, 1989; Pfeffermann, 1993). Hence, unbiased active learning allows for oversampling of the most informative instances without compromising unbiasedness.

An unbiased active learning method is presented in Algorithm 4.2. The algorithm sequentially samples new training examples from the pool of available instances. In each step, n_k instances are selected at random according to a

probability sampling design, and the corresponding labels y_i are retrieved. The prediction model is then updated according to (4.2) for some loss function $\ell(\boldsymbol{\theta}; \boldsymbol{x}, y)$. We assume in Algorithm 4.2 that sampling is conducted according to a multinomial sampling design, but note that other sampling designs may also be considered. The algorithm continues until a pre-specified maximal number of iterations is reached, or some target on the predictive performance of the model is satisfied.

Algorithm 4.2. Unbiased active learning.

INPUT: Sampling frame $\mathcal{D}$, number of iterations K, batch sizes $\{n_k\}_{k=1}^K$.
INITIALISATION: Let $m_0 = 0$, $\mathcal{L}_0 = \varnothing$.

1: **for** k = 1, . . . , K **do**
2: Calculate sampling scheme $\boldsymbol{\mu}_k$.
3: Draw vector $\boldsymbol{s}_k = (s_{k1}, \ldots, s_{kN}) \sim \text{Multinomial}(n_k, \boldsymbol{\mu}_k / n_k)$.
4: Retrieve the value(s) of y_i for the selected instance(s).
5: Let $\mathcal{L}_k = \mathcal{L}_{k-1} \cup \{i \in \mathcal{D} : s_{ki} > 0\}$ and $m_k = m_{k-1} + n_k$.
6: Update the sampling weights

$$w_{ki} = \frac{1}{m_k} \left(m_{k-1} w_{k-1,i} + n_k \frac{s_{ki}}{\mu_{ki}} \right), \quad i \in \mathcal{D}.$$

7: Update the parameter

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}^{(k)} = \arg\min_{\boldsymbol{\theta}} \sum_{i \in \mathcal{L}_k} w_{ki} \ell(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i).$$

8: STOP if performance target is reached. ELSE continue.
9: **end for**

For a generalised linear model (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989), the optimal sampling scheme for a multinomial sampling design is in Paper I shown to be

$$\mu_{ki} = n_k \frac{\sqrt{h_{ii}(\boldsymbol{\theta}_0)}}{\sum_{j \in \mathcal{D}} \sqrt{h_{jj}(\boldsymbol{\theta}_0)}}, \tag{4.3}$$

where $h_{ii}(\boldsymbol{\theta}_0) = \text{Var}_{Y_i \sim f_{\boldsymbol{\theta}_0}(y|\boldsymbol{x}_i)}(Y_i) \boldsymbol{x}_i^\mathsf{T} \boldsymbol{H}(\boldsymbol{\theta}_0)^{-1} \boldsymbol{x}_i$ is the statistical leverage score pertaining to instance i (Hoaglin and Welsch, 1978; Pregibon, 1981). Note that we have introduced a random variable Y_i to capture our uncertainty in the unknown value of the outcome y_i. The sampling scheme (4.3) is optimal in the sense of minimising the anticipated generalisation error of the estimator with respect to the full-data empirical risk. In practice the optimal parameter $\boldsymbol{\theta}_0$ is

unknown and the leverage score $h_{ii}(\boldsymbol{\theta}_0)$ evaluated at the current estimate $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}^{(k)}$. A similar result was also obtained by Zhang et al. (2021).

It was noted in a simultaneous publication by Ma et al. (2020) on optimal subsampling for linear regression, that the variance of the residual around the fitted value in a (generalised) linear model should be shrunk by a factor $1 - h_{ii}(\boldsymbol{\theta}_0)$. This is due to the influence the data point exerts on its own prediction (cf. Hoaglin and Welsch, 1978; Pregibon, 1981). With this modification, the corresponding optimal sampling scheme becomes

$$\mu_{ki} = n_k \frac{\sqrt{h_{ii}(1 - h_{ii})}}{\sum_{j \in \mathcal{D}} \sqrt{h_{jj}(1 - h_{jj})}}, \tag{4.4}$$

where, as a above $h_{ii} = h_{ii}(\boldsymbol{\theta}_0)$ may depend on the optimal parameter $\boldsymbol{\theta}_0$. The difference between (4.3) and (4.4) is of limited practical importance when effective number of parameters is small in relation to the full data size N. Indeed, it holds that $\sum_i h_{ii} = p$, where p is the dimension of the parameter vector (Hoaglin and Welsch, 1978), and consequently that $h_{ii} = O(p/N)$ and $h_{ii}(1 - h_{ii}) \approx h_{ii}$ for essentially all instances $i \in \mathcal{D}$ when $p \ll N$.

4.3 Active sampling

In this section we describe an active sampling method for finite population inference with optimal subsamples. The method is summarised in Algorithm 4.3. We consider a scalar finite population characteristic $\tau = h(\boldsymbol{t}_{\boldsymbol{y}})$, for some differentiable function h and vector total $\boldsymbol{t}_{\boldsymbol{y}}$. The algorithm proceeds in K iterations $k = 1, \ldots, K$ and chooses, in each iteration, n_k new instances at random from $\mathcal{D}$. We assume here that sampling is conducted according to a multinomial design, although other sampling designs may also be considered. Once a new batch of instances has been selected we observe the corresponding data $\boldsymbol{y}_i$ and update our estimate of the characteristic of interest. The process continues until a pre-specified maximal number of iterations K is reached, or the target characteristic is estimated with sufficient precision. Methods for variance estimation, needed to assess the precision of the estimator, are described in Paper III.

A key component of the active sampling algorithm is the inclusion of an auxiliary model $f(\boldsymbol{y}_i|\boldsymbol{z}_i)$ for the distribution of the unobserved data $\boldsymbol{y}_i$ given auxiliary variables $\boldsymbol{z}_i$. At this stage, any prediction model or machine learning algorithm can be used. By gathering data in a sequential manner, we may iteratively update our predictions on yet unseen data. Doing so, we are able to learn from past observations how to sample in an optimal way in future

iterations. The better the model is in predicting the values of the unobserved data, the greater the potential benefit of optimal subsampling.

Under the multinomial sampling design, the optimal sampling scheme for estimating a scalar finite population characteristic $\tau = h(\boldsymbol{t}_{\boldsymbol{y}})$ is in Paper III shown to be on the form

$$\mu_{ki} = n_k \frac{\sqrt{c_i}}{\sum_{j \in \mathcal{D}} \sqrt{c_j}}. \tag{4.5}$$

Here c_i is a function of the gradient $\nabla \boldsymbol{h}(\boldsymbol{u})|_{\boldsymbol{u}=\boldsymbol{t}_{\boldsymbol{y}}}$, prediction $\hat{\boldsymbol{y}}_i$, and residual covariance matrix $\boldsymbol{\Sigma}_i$. Explicit expressions of the optimal coefficients c_i are presented in Table 4.1 for some common characteristics, including the finite population total, mean, and ratio. The sampling schemes in Table 4.1 are optimal in the sense of minimising the anticipated variance of the estimator for $h(\boldsymbol{t}_{\boldsymbol{y}})$ under the auxiliary model $f(\boldsymbol{y}|\boldsymbol{z})$. Similar procedures may be employed to derive a active sampling methods for vector-valued characteristics and general optimality criteria.

Algorithm 4.3. Active sampling.

INPUT: Sampling frame $\mathcal{D}$, precision target δ, maximal number of iterations K, batch sizes $\{n_k\}_{k=1}^{K}$.

INITIALISATION: Let $m_0 = 0$, $\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(0)} = \boldsymbol{0}$.

1: **for** k = 1, . . . , K **do**
2: If $k > 1$: Train prediction model $f(\boldsymbol{y}_i|\boldsymbol{z}_i)$ on the labelled dataset.
3: Calculate sampling scheme $\boldsymbol{\mu}_k$ (see Table 4.1).
4: Draw vector $\boldsymbol{s}_k = (s_{k1}, \ldots, s_{kN}) \sim \text{Multinomial}(n_k, \boldsymbol{\mu}_k/n_k)$.
5: Retrieve data $\boldsymbol{y}_i$ for the selected instance(s) $i \in \mathcal{L}_k := \{i \in \mathcal{D} : s_{ki} > 0\}$.
6: Let $m_k = m_{k-1} + n_k$,

$$\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)} = m_k^{-1} \left(m_{k-1} \hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k-1)} + n_k \sum_{i \in \mathcal{L}_k} \mu_{ki}^{-1} s_{ki} \boldsymbol{y}_i \right), \quad \hat{\tau}^{(k)} = h(\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)}).$$

7: **if** $\widehat{\text{Var}}(\hat{\tau}^{(k)}) < \delta$ **then**
8: STOP. Sufficient precision is reached.
9: **end if**
10: **end for**

Table 4.1. Coefficients c_i for the optimal sampling scheme (4.5) to estimate a finite population characteristic $\tau = h(\boldsymbol{t_y})$. Hats represent estimates or predicted values. $\sigma_{x,i}, \sigma_{y,i}$ are residual variances, $\rho_{xy,i}$ the residual correlation, and $\boldsymbol{\Sigma}_i$ the residual covariance matrix of a random vector $\boldsymbol{Y}_i$ with density $f(\boldsymbol{y}_i|\boldsymbol{z}_i)$. In the active sampling framework, all estimates and predictions are updated iteratively according to Algorithm 4.3.

Target characteristic	c_i
Total t_y	$\hat{y}_i^2 + \sigma_{y,i}^2$
Mean t_y/N (Horvitz-Thompson estimator $\hat{t}_y/N$)	$\hat{y}_i^2 + \sigma_{y,i}^2$
Mean t_y/N (Hájek estimator $\hat{t}_y/\hat{N}$)	$(\hat{y}_i - \hat{t}_y/\hat{N})^2 + \sigma_{y,i}^2$
Ratio $r_{yx} = t_y/t_x$	$(\hat{y}_i - \hat{r}_{yx}\hat{x}_i)^2 + \sigma_{y,i}^2 + \hat{r}_{yx}^2\sigma_{x,i}^2 - 2\hat{r}_{yx}\rho_{xy,i}\sigma_{x,i}\sigma_{y,i}$
General $h(\boldsymbol{t_y})$	$(\nabla h(\boldsymbol{u})^\mathsf{T}\hat{\boldsymbol{y}}_i)^2\big\|_{\boldsymbol{u}=\hat{\boldsymbol{t}}_{\boldsymbol{y}}} + \nabla h(\boldsymbol{u})^\mathsf{T}\boldsymbol{\Sigma}_i\nabla h(\boldsymbol{u})\big\|_{\boldsymbol{u}=\hat{\boldsymbol{t}}_{\boldsymbol{y}}}$

4.4 A martingale central limit theorem

In this section we establish the asymptotic properties of the active sampling estimator for a finite population total. A generalisation to multivariate estimators and to characteristics defined as smooth functions of totals may be found in the appendix of Paper III. For a sequence of random variables $\{X, X_n, n \geq 1\}$, we let $X_n \xrightarrow{d} X$ and $X_n \xrightarrow{p} X$ denote convergence of X_n to X in distribution and in probability, respectively.

Consider the setup of Chapter 4.1. Let $t_y = \sum_{i\in\mathcal{D}} y_i$ be the total of interest, $\hat{t}_{y,k} = \sum_{i\in\mathcal{D}} \mu_{ki}^{-1} S_{ki} y_i$ the Hansen-Hurwitz estimator from the subsample obtained at stage k, and $\hat{t}_y^{(K)} = m_K^{-1}\sum_{k=1}^K n_k \hat{t}_{y,k}$ the pooled estimator from the first K subsamples. Let $\sigma_k^2 = \mathrm{Var}(\hat{t}_{y,k}|\boldsymbol{S}_1, \ldots, \boldsymbol{S}_{k-1})$, $A_K^2 = \sum_{k=1}^K n_k^2\sigma_k^2$ and $b_K^2 = \mathrm{Var}(\sum_{k=1}^K n_k \hat{t}_{y,k})$. Also assume that:

(A1) S_{ki}/μ_{ki} have uniformly bounded second moments,

(A2) $b_K \to \infty$ as $K \to \infty$,

(A3) $A_K^2 b_K^{-2} \xrightarrow{p} 1$ as $K \to \infty$, and

(A4) there exists a collection of random variables $\{\hat{\sigma}_k^2\}_{k=1}^K$ such that $\mathrm{E}[\hat{\sigma}_k^2|\boldsymbol{S}_1, \ldots, \boldsymbol{S}_{k-1}] = \sigma_k^2$ and $b_K^{-2}\mathrm{Var}(\sum_{k=1}^K n_k^2\hat{\sigma}_k^2)$ are uniformly bounded.

Using the martingale central limit theorem of Brown (1971), we obtain the following result:

Proposition 4.1 (Martingale central limit theorem)**.**
Assume that (A1)–(A3) holds. Then

$$\frac{\hat{t}_y^{(K)} - t_y}{b_K/m_K} \xrightarrow{d} \mathcal{N}(0,1) \quad \textit{and} \tag{4.6}$$

$$b_K^{-2} \sum_{k=1}^{K} n_k^2 \left(\hat{t}_{y,k} - \hat{t}_y^{(K)}\right)^2 \xrightarrow{p} 1 \quad \textit{as } K \to \infty. \tag{4.7}$$

If (A4) also holds, then

$$b_K^{-2} \sum_{k=1}^{K} n_k^2 \hat{\sigma}_k^2 \xrightarrow{p} 1 \quad \textit{as } K \to \infty. \tag{4.8}$$

The first result (4.6) establishes consistency and asymptotic normality of the active sampling estimator $\hat{t}_y^{(K)}$ of the finite population total t_y. The second result (4.7) proves the consistency of the martingale variance estimator $\widehat{\mathrm{Var}}(\hat{t}_y^{(K)}) = m_K^{-2} \sum_{k=1}^{K} n_k^2 (\hat{t}_{y,k} - \hat{t}_y^{(K)})^2$ for the variance of $\hat{t}_y^{(K)}$. The third result (4.8) proves the consistency of the pooled variance estimator $\widetilde{\mathrm{Var}}(\hat{t}_y^{(K)}) = m_K^{-2} \sum_{k=1}^{K} n_k^2 \hat{\sigma}_k^2$. Unbiased estimators $\hat{\sigma}_k^2$ for the conditional variances σ_k^2 may be obtained using the methods of Chapter 2.1.3.

The first assumption (A1) is fulfilled when the mean inclusion variables μ_{ki} are properly bounded away from zero. The second assumption (A2) requires the total variance $\mathrm{Var}(\sum_{k=1}^{K} n_k \hat{t}_{y,k})$ to tend to infinity with K. This is a plausible assumption in most realistic applications since the precision of the individual estimators $\hat{t}_{y,k}$ usually are of order $O(n_k)$ and n_k are bounded. The third assumption (A3) states that the sum of conditional variances should asymptotically behave like the total variance. Hence, the statistical properties of the active sampling estimator can be deduced from a single realisation of the sequence $\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots$. The plausibility of this assumption may be ensured by designing active sampling strategies so that each individual observation has a limited influence on the selections in future iterations. Assumptions on the form of (A1)–(A3) are frequent in the martingale literature (cf. Brown, 1971; McLeish, 1974; Helland, 1982). The fourth assumption (A4) is related to the consistency of variance estimators in survey sampling. This assumption is generally fulfilled when S_{ki}/μ_{ki} have uniformly bounded forth moments, and the joint expectations of the selection variables S_{ki} at each sampling stage k are properly bounded away from zero (cf. Fuller, 2009).

5 Summary of papers

5.1 Paper I

Introduction. We consider a statistical learning problem of fitting a model $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ to a subset of a random sample $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, with the aim to accurately predict the outcomes y_i given observed inputs $\boldsymbol{x}_i$. The predictors $\boldsymbol{x}_i$ are known for all members of the initial sample, but the outcomes or labels y_i are affordable to observe only for a smaller subset. The question arises: how should the training dataset be chosen (i.e., which of the y_i's should be observed) for optimal predictive performance?

Contributions. We consider an unbiased active learning algorithm based on unequal probability sampling and inverse probability weighting. We conduct an asymptotic analysis of the generalisation error and derive corresponding optimal sampling schemes. The resulting sampling schemes depend both on the prediction uncertainty and on the influence on model fitting through the location of data points in the feature space. A connection to leverage sampling is established, revealing that influential instances should be oversampled for optimal predictive performance.

Experiments. We evaluate the empirical performance of the proposed method for binary classification on six benchmark data sets. For reference, we also include simple random sampling, probabilistic uncertainty sampling, and deterministic uncertainty sampling.

Results. The unbiased active learning algorithm targeting the generalisation error overall achieves the best performance in terms of the negative log-likelihood of the predictions. A sample size reduction of up to 23% is achieved

compared to simple random sampling for the same level of performance. In contrast, probabilistic uncertainty sampling performs worse than simple random sampling with respect to essentially all performance metrics on four of the datasets. Deterministic uncertainty sampling produces poorly calibrated predictions with a severe bias towards the majority class in two of the examples, and overfitted class probability estimates on the majority of the examples.

5.2 Paper II

Introduction. Naturalistic driving studies generate tremendous amounts of traffic data. For instance, the SHRP2 project collected more than a million hours of driving data, including video and recordings of vehicle kinematics etc. The great cost associated with video annotation implies that statistical analyses based on video data must be restricted to only a limited subset of the original database. Choosing this subset in a manner that captures as much information as possible is essential.

Contributions. We derive optimal sampling schemes for a weighted maximum likelihood estimator with respect to the L-optimality criterion, under a Poisson sampling design. We show how auxiliary information may be used to implement optimal subsampling methods in practice.

Experiments. The methodology is illustrated using data collected in Sweden as part of the large-scale European field operational test (euroFOT). We describe how automatic measurements, readily available in a naturalistic driving database, may be utilised for selection of time segments for annotation that are the most informative with regards to detection of potential associations between driving behaviour and a consecutive safety critical event.

Results. Poisson sampling optimised for a specific linear combination of parameters generally results in an increased precision of the corresponding estimates. A variance reduction by 21–48% with optimal subsampling compared to simple random sampling is demonstrated when good auxiliary information is available. On the other hand, application of the method to variables with poor auxiliary information may result in increased variance of the estimator.

5.3 Paper III

Introduction. We consider the problem of estimating a simple finite population characteristic, such as a total, mean, or ratio. This classical problem has received much attention in the survey sampling literature. However, the possibilities offered by machine learning in this context have not yet been fully explored.

Contributions. We develop an active sampling framework for estimating a finite population characteristic $\theta = h(\boldsymbol{t_y})$, for a differentiable function h and vector of finite population totals $\boldsymbol{t_y}$. The active sampling algorithm iterates between estimation and data collection with optimal subsamples, guided by machine learning predictions on yet unseen data. Using a martingale central limit theorem, we establish consistency and asymptotic normality of the active sampling estimator for θ under mild assumptions. Methods for variance estimation are proposed and consistency of the variance estimators is proven.

Experiments. The active sampling method is implemented for an application in scenario generation for virtual safety assessment of an advanced driver assistance system. The dataset consists of 44,220 observations generated through variations of 44 reconstructed real rear-end crashes.

Results. Our theoretical results are confirmed empirically in our experiments. The asymptotic theory and suggested variance estimators produce confidence intervals with coverage rates that reach the nominal confidence level already at small to moderate sample sizes. Substantial improvements over traditional importance sampling methods are demonstrated, with sample size reductions of 7–48% for the same level of performance in terms of the root mean squared error of the estimator.

5.4 Paper IV

Introduction. We consider a subsampling estimator

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \frac{S_i}{\mu_i} \ell_i(\boldsymbol{\theta})$$

for a characteristic $\boldsymbol{\theta}_0$ defined by $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta})$. Here S_i is the number of times an element i is selected by the sampling mechanism, μ_i the corresponding expected number of selections, and $\ell_i(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}; \boldsymbol{v}_i)$ a loss function describing the loss associated with the parameter value $\boldsymbol{\theta}$ given a data vector $\boldsymbol{v}_i$.

Contributions. We present a theory of optimal design for general data subsampling problems. Optimality conditions are derived for a general class of estimators, sampling designs, and optimality criteria. A fixed-point iteration method is suggested for finding an optimal sampling scheme under a Poisson or multinomial sampling design. We also study optimal design from a distance-minimising perspective, and establish equivalence to traditional optimality criteria. This naturally leads us to a novel class of linear optimality criteria with good theoretical and practical properties, including computational tractability and invariance under non-singular affine transformations of the data and under a re-parameterisation of the model.

Experiments. The methodology is evaluated on the dataset from Paper III for parametric density estimation, regression modelling, and inference for a finite population vector characteristic.

Results. A D-optimal sampling design is found within a few fixed-point iterations with the suggested algorithm. The proposed invariant linear optimality criteria achieve 92–99% D-efficiency with 90–95% lower computational demand. In contrast, the A-optimality criterion has only 46% and 60% D-efficiency on two of the examples.

6 Discussion

Optimal subsampling methods have seen a huge increase in popularity over the past few years for analysis of massive datasets and measurement-constrained experiments. This thesis contributes to this development by presenting a framework for optimal design in general subsampling problems using auxiliary information and sequential optimal design.

Using a martingale central limit theorem, under mild assumptions we establish consistency and asymptotic normality of the active sampling estimator of a simple finite population characteristic. Combining martingale limit theory (Hall and Heyde, 1980) with the asymptotics of estimating equation estimators in survey sampling (Binder, 1983), consistency and asymptotic normality of more complex estimators may be deduced by similar means (cf. Zhang et al., 2021; Wang et al., 2022). We conjecture that a similar result holds also in the case when the number of sample stages is fixed and the subsample sizes tend to infinity. A thorough treatment of this issue is a possible topic for further research.

A limitation of the presented methodology is the underlying assumption of a smooth loss function and parameter vector of fixed dimension $p \ll N$. Although similar subsampling methods may be employed also for non-regular problems, e.g., in high-dimensional settings and for non-smooth loss functions, the asymptotic properties and optimal subsampling methods have not been established in these regimes. It would be interesting to study optimal subsampling methods in high-dimensional and non-parametric settings, for instance for kernel generalised linear models (Zhu and Hastie, 2005; Cawley et al., 2007; Hofmann et al., 2008).

A potential drawback of the design-based inference framework adopted in this thesis is the strong influence of the sampling weights on estimation. This may cause unstable performance and loss of efficiency, in particular when the subsample is optimised for one specific aim but later used for other purposes,

as often is the case in practice. Consequently, some authors have suggested unweighted methods for more efficient estimation (Ma et al., 2015; Wang, 2019; Wang et al., 2022). There is also a vast amount of publications in the survey sampling literature on smoothing, trimming and calibration of the sampling weights to improve efficiency (see, e.g., Deville and Särndal, 1992; Kott, 2016; Haziza and Beaumont, 2017). Another possibility to improve estimator efficiency is to utilise auxiliary information in the estimation stage. Such methods are prevalent across many sub-fields of statistics, including generalised regression estimation in survey sampling (Cassel et al., 1976; Särndal et al., 2003; Ta et al., 2020), control variates in the Monte Carlo literature (Fishman, 1996; Quiroz et al., 2021), and augmented inverse probability weighting estimators in causal inference and semi-parametric inference with missing data (Robins et al., 1994; Robins, 1999; Scharfstein et al., 1999). Although there is already an extensive literature on these methods, their use for analysis massive datasets and measurement-constrained experiments in the general setting considered in this thesis has to our knowledge not been widely employed. Further studies in this direction are encouraged.

References

Ai, M., Wang, F., Yu, J., and Zhang, H. (2021a). Optimal subsampling for large-scale quantile regression. *Journal of Complexity*, 62:101512.

Ai, M., Yu, J., Zhang, H., and Wang, H. (2021b). Optimal subsampling algorithms for big data regressions. *Statistica Sinica*, 31:749–772.

Anderson, R., Doecke, S., Mackenzie, J., and Ponte, G. (2013). Potential benefits of autonomous emergency braking based on in-depth crash reconstruction and simulation. In *Proceedings of the 23rd International Conference on Enhanced Safety of Vehicles*.

Atkinson, A. C. and Donev, A. N. (1992). *Optimum Experimental Designs*. Clarendon Press, Oxford.

Bach, F. R. (2007). Active learning for misspecified generalized linear models. In *Advances in Neural Information Processing Systems 19*.

Beygelzimer, A., Dasgupta, S., and Langford, J. (2009). Importance weighted active learning. In *Proceedings of the 26th International Conference on Machine Learning*.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51(3):279–292.

Box, G. E. P. and Hunter, W. G. (1965). Sequential design of experiments for nonlinear models. In *Proceedings of the IBM Scientific Computing Symposium on Statistics*.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

Brown, B. M. (1971). Martingale central limit theorems. *The Annals of Mathematical Statistics*, 42(1):59–66.

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620.

Cawley, G. C., Janacek, G. J., and Talbot, N. L. C. (2007). Generalised kernel machines. In *2007 International Joint Conference on Neural Networks*.

Chen, M., Hao, Y., Hwang, K., Wang, L., and Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*, 5:8869–8879.

Chu, W., Zinkevich, M., Li, L., Thomas, A., and Tseng, B. (2011). Unbiased online active learning in data streams. In *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Cohn, D. A. (1996). Neural network exploration using optimal experiment design. *Neural Networks*, 9(6):1071–1083.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

Deldossi, L. and Tommasi, C. (2022). Optimal design subsampling from big datasets. *Journal of Quality Technology*, 54(1):93–101.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M. A., Buchanan-King, M., and Hankey, J. M. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences*, 113(10):2636–2641.

Drovandi, C. C., Holmes, C. C., McGree, J. M., Mengersen, K., Richardson, S., and Ryan, E. G. (2017). Principles of experimental design for big data analysis. *Statistical Science*, 32(3):385–404.

Farquhar, S., Gal, Y., and Rainforth, T. (2021). On statistical bias in active learning: How and when to fix it. In *International Conference on Learning Representations*.

Fishman, G. S. (1996). *Monte Carlo*. Springer, New York.

Ford, I. and Silvey, S. D. (1980). A sequentially constructed design for estimating a nonlinear parametric function. *Biometrika*, 67(2):381–388.

Fuller, W. A. (2009). *Sampling Statistics*. Wiley, Hoboken.

Ganti, R. and Gray, A. (2012). UPAL: Unbiased pool based active learning. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*.

Geuten, K., Massingham, T., Darius, P., Smets, E., and Goldman, N. (2007). Experimental design criteria in phylogenetics: Where to add taxa. *Systematic biology*, 56(4):609–622.

Grafström, A. (2010). *On Unequal Probability Sampling Designs*. PhD thesis, Department of Mathematics and Mathematical Statistics, Umeå University.

Hájek, J. (1981). *Sampling from a Finite Population*. Marcel Dekker, New York.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, New York.

Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4):333–362.

Hartley, H. O. and Sielken, R. L. (1975). A "super-population viewpoint" for finite population sampling. *Biometrics*, 31(2):411–422.

Haussler, D. (1989). Learning conjunctive concepts in structural domains. *Machine Learning*, 4(1):7–40.

Haziza, D. and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32(2):206–226.

Helland, I. S. (1982). Central limit theorems for martingales with discrete or continuous time. *Scandinavian Journal of Statistics*, 9(2):79–94.

Hoaglin, D. C. and Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician*, 32(1):17–22.

Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96.

Korn, E. L. and Graubard, B. I. (1995). Analysis of large health surveys: Accounting for the sampling design. *Journal of the Royal Statistical Society. Series A*, 158(2):263–295.

Kossen, J., Farquhar, S., Gal, Y., and Rainforth, T. (2022). Active surrogate estimators: An active learning approach to label-efficient model evaluation. In *Advances in Neural Information Processing Systems*.

Kott, P. S. (2016). Calibration weighting in survey sampling. *WIREs Computational Statistics*, 8(1):39–53.

Landsman, V. and Graubard, B. I. (2013). Efficient analysis of case-control studies with sample weights. *Statistics in Medicine*, 32(2):347–360.

Lewis, D. D. and Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning. In *Machine Learning: Proceedings of the Eleventh International Conference*.

Lewis, D. D. and Gale, W. A. (1994). A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Ma, P., Mahoney, M. W., and Yu, B. (2015). A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 16(27):861–911.

Ma, P., Zhang, X., Xing, X., Ma, J., and Mahoney, M. W. (2020). Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*.

MacKay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604.

Magnus, J. R. and Neudecker, H. (2019). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, Hoboken.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. CRC Press, Boca Raton.

McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *The Annals of Probability*, 2(4):620–628.

Meng, C., Xie, R., Mandal, A., Zhang, X., Zhong, W., and Ma, P. (2021). LowCon: A design-based subsampling approach in a misspecified linear model. *Journal of Computational and Graphical Statistics*, 30(3):694–708.

Mullins, G. E., Stankiewicz, P. G., Hawthorne, R. C., and Gupta, S. K. (2018). Adaptive generation of challenging scenarios for testing and evaluation of autonomous vehicles. *Journal of Systems and Software*, 137:197–215.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A*, 135(3):370–384.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201):101–116.

Petersen, K. B. and Pedersen, M. S. (2012). *The Matrix Cookbook*. Version 20121115. Technical University of Denmark.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61(2):317–337.

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā: The Indian Journal of Statistics, Series B*, 61(1):166–186.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, 9(4):705–724.

Pronzato, L. (2006). On the sequential construction of optimum bounded designs. *Journal of Statistical Planning and Inference*, 136(8):2783–2804.

Pronzato, L. and Pázman, A. (2013). *Design of Experiments in Nonlinear Models*. Springer, New York.

Pukelsheim, F. (1993). *Optimal Design of Experiments*. Wiley, New York.

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D., editors (2008). *Dataset Shift in Machine Learning*. MIT Press, Cambridge.

Quiroz, M., Tran, M.-N., Villani, M., Kohn, R., and Dang, K.-D. (2021). The block-Poisson estimator for optimally tuned exact subsampling MCMC. *Journal of Computational and Graphical Statistics*, 30(4):877–888.

Righi, A. (2019). Assessing migration through social media: a review. *Mathematical Population Studies*, 26(2):80–91.

Robins, J. M. (1999). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin-Bleuer, S. and Kratina, I. S. (2005). On the two-phase framework for joint model and design-based inference. *The Annals of Statistics*, 33(6):2789–2810.

Särndal, C. E., Swensson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer, New York.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Schein, A. I. and Ungar, L. H. (2007). Active learning for logistic regression: An evaluation. *Machine Learning*, 68(3):235–265.

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5:119–127.

Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics*. Chapman & Hall, Boca Raton.

Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114.

Seyedi, M., Koloushani, M., Jung, S., and Vanli, A. (2021). Safety assessment and a parametric study of forward collision-avoidance assist based on real-world crash simulations. *Journal of Advanced Transportation*. Advance online publication. doi: 10.1155/2021/4430730.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244.

Silvey, S. (1980). *Optimal Design*. Chapman & Hall, London.

Sirpitzi, R. E., Miller, F., and Burman, C. F. (2023). Robust optimal designs using a model misspecification term. *Metrika*. Advance online publication. doi: 10.1007/s00184-023-00893-6.

Skinner, C. J. (1989). Domain Means, Regression and Multivariate Analysis. In Skinner, C. J., Holt, D., and Smith, T. M. F., editors, *Analysis of Complex Surveys*, pages 80–87. Wiley, Chichester.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician*, 56(1):29–38.

Sugiyama, M. (2006). Active learning in approximately linear regression based on conditional expectation of generalization error. *Journal of Machine Learning Research*, 7(6):141–166.

Sugiyama, M. and Nakajima, S. (2009). Pool-based active learning in approximate linear regression. *Machine Learning*, 75(3):249–274.

Sun, J., Zhou, H., Xi, H., Zhang, H., and Tian, Y. (2022). Adaptive design of experiments for safety evaluation of automated vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):14497–14508.

Ta, T., Shao, J., Li, Q., and Wang, L. (2020). Generalized regression estimators with high-dimensional covariates. *Statistica Sinica*, 30(3):1135–1154.

Tillé, Y. (2006). *Sampling Algorithms*. Springer, New York.

Tommasi, C. (2012). Optimal design robust to a misspecified model. *Communications in Statistics - Simulation and Computation*, 41(7):1220–1231.

van Schagen, I. and Sagberg, F. (2012). The potential benefits of naturalistic driving for road safety research: Theoretical and empirical considerations and challenges for the future. *Procedia - Social and Behavioral Sciences*, 48:692–701.

Vapnik, V. (1991). Principles of risk minimization for learning theory. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*.

Wand, M. P. (2002). Vector differential calculus in statistics. *The American Statistician*, 56(1):55–62.

Wang, H. (2019). More efficient estimation for logistic regression with optimal subsamples. *Journal of Machine Learning Research*, 20(132):1–59.

Wang, H. and Ma, Y. (2021). Optimal subsampling for quantile regression in big data. *Biometrika*, 108(1):99–112.

Wang, H., Yang, M., and Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114(525):393–405.

Wang, H., Zhu, R., and Ma, P. (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844.

Wang, J., Wang, H., and Xiong, S. (2022). Unweighted estimation based on optimal sample under measurement constraints. *Canadian Journal of Statistics*. Advance online publication. doi: 10.1002/cjs.11753.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61(3):439–447.

Welch, W. J. (1984). Computer-aided design of experiments for response estimation. *Technometrics*, 26(3):217–224.

Winkelbauer, M., Eichhorn, A., Sagberg, F., and Backer-Grøndahl, A. (2010). Naturalistic driving. In *Data and Mobility: Transforming Information into Intelligent Traffic and Transportation Services. Proceedings of the Lakeside Conference 2010*.

Wynn, H. P. (1982). Optimum submeasures with application to finite population sampling. In Gupta, S. S. and Berger, J. O., editors, *Statistical Decision Theory and Related Topics III*, pages 485–495. Academic Press, New York.

Yao, Y. and Wang, H. (2019). Optimal subsampling for softmax regression. *Statistical Papers*, 60:585–599.

Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society. Series B*, 15(2):253–261.

Yu, J., Wang, H., Ai, M., and Zhang, H. (2022). Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association*, 117(537):265–276.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-First International Conference on Machine Learning*.

Zhan, X., Wang, Y., and Chan, A. B. (2022). Asymptotic optimality for active learning processes. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*.

Zhang, T., Ning, Y., and Ruppert, D. (2021). Optimal sampling for generalized linear models under measurement constraints. *Journal of Computational and Graphical Statistics*, 30(1):106–114.

Zhu, J. and Hastie, T. (2005). Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, 14(1):185–205.

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., and Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing*, 74:26–39.