OPTIMAL SUBSAMPLING DESIGNS

Henrik Imberg
imbergh@chalmers.se

Marina Axelson-Fisk
marina.axelson-fisk@chalmers.se

Johan Jonasson
jonasson@chalmers.se

Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg, Sweden

ABSTRACT

Subsampling is commonly used to overcome computational and economical bottlenecks in the analysis of finite populations and massive datasets. Existing methods are often limited in scope and 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. We present a theory of optimal design for general data subsampling problems, including finite population inference, parametric density estimation, and regression modelling. Our theory encompasses and generalises most existing methods in the field of optimal subdata selection based on unequal probability sampling and inverse probability weighting. We derive optimality conditions for a general class of optimality criteria, and present corresponding algorithms for finding optimal sampling schemes under Poisson and multinomial sampling designs. We present a novel class of transformation- and parameterisation-invariant linear optimality criteria which enjoy the best of two worlds: the computational tractability of A-optimality and invariance properties similar to D-optimality. The methodology is illustrated on an application in the traffic safety domain. In our experiments, 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.

Keywords A-optimality · D-optimality · L-optimality · M-estimation · inverse probability weighting · unequal probability sampling.

1 Introduction

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

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta} \in \boldsymbol{\Omega}}{\arg\min}\, \ell_0(\boldsymbol{\theta}), \tag{1}$$

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

$$\ell_0(\boldsymbol{\theta}) = \sum_{i \in \mathcal{D}} \ell_i(\boldsymbol{\theta}), \quad \ell_i(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}; \boldsymbol{v}_i), \tag{2}$$

with summation over some index set $\mathcal{D} = \{1, \ldots, N\}$, where $\boldsymbol{v}_i$ is a data vector associated with a member $i \in \mathcal{D}$. Under these assumptions, $\boldsymbol{\theta}_0$ may also be defined as the unique solution to the estimation equation

$$\sum_{i \in \mathcal{D}} \boldsymbol{\psi}_i(\boldsymbol{\theta}) = \mathbf{0}, \quad \boldsymbol{\psi}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{v}_i). \tag{3}$$

The data is on the form $\boldsymbol{v}_i = \boldsymbol{y}_i \in \mathcal{Y}$ or $\boldsymbol{v}_i = (\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{X} \times \mathcal{Y}$, where $\boldsymbol{y}_i$ is a response vector and $\boldsymbol{x}_i$ a vector of explanatory variables. We will generally not distinguish between the case with and without explanatory variables, and throughout we write the data as $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, keeping in mind that the first entry may be null and $\mathcal{X}$ the empty set. One may interpret (1)–(2) as an empirical risk minimisation problem (Vapnik, 1991). Hence, we will refer to $\ell_0(\boldsymbol{\theta})$ as the (full-data) empirical risk and to $\boldsymbol{\theta}_0$ as the (full-data) empirical risk minimiser (ERM).

The setting above 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, which will be considered further in the Application and Examples in Section 6, include:

i) Finite population inference: 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 $\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y}_i$ may be written on the form (1)–(3) 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 $y_1, \dots, y_N$ from a probability distribution with density function $f_{\boldsymbol{\theta}}(y)$, the maximum likelihood estimate of $\boldsymbol{\theta}$ may be written on the form (1)–(3) 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 (1)–(3) define an estimate of the vector of regression coefficients $\boldsymbol{\theta}$.

Now consider a situation where inference based on the full data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i\in\mathcal{D}}$ is prohibited by economic or computational constraints. For instance, the index set may be so large that complete enumeration to observe the full data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i\in\mathcal{D}}$ is practically or economically unfeasible. This is the typical situation in finite population inference (Neyman, 1938; Hansen and Hurwitz, 1943; Horvitz and Thompson, 1952). Some variables may be expensive to measure and hence affordable to observe only for a small number of instances $i \in \mathcal{D}$, a situation known as a measurement-constrained experiment (Wang et al., 2017; Meng et al., 2021; Zhang et al., 2021; Imberg et al., 2022b). Another example is when the full data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i\in\mathcal{D}}$ is available, but the size N of the dataset is so large that estimation of $\boldsymbol{\theta}$ using (1)–(2) is computationally unfeasible (Ma et al., 2015; Drovandi et al., 2017; Wang et al., 2018; Deldossi and Tommasi, 2022; Dai et al., 2022). In either case, we may search for an approximate solution based on a subset $\mathcal{S} \subset \mathcal{D}$ of size $n \ll N$.

In this paper we focus on methods based on data subsampling through unequal probability sampling and inverse probability weighting. Specifically, we consider an estimator of the form

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Omega}}{\arg\min}\, \hat{\ell}_{\boldsymbol{\mu}}(\boldsymbol{\theta}), \tag{4}$$

$$\hat{\ell}_{\boldsymbol{\mu}}(\boldsymbol{\theta}) = \sum_{i\in\mathcal{S}} \frac{S_i}{\mu_i}\ell_i(\boldsymbol{\theta}), \tag{5}$$

where S_i is the number of times an element $i \in \mathcal{D}$ is selected by the sampling mechanism, μ_i the corresponding expected number of selections, and $\mathcal{S} = \{i \in \mathcal{D} : S_i > 0\}$ the random set of selected elements. One may recognise (5) as the Hansen-Hurwitz estimator (Hansen and Hurwitz, 1943) of the full-data empirical risk function (2). Hence, we refer to $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ as the Hansen-Hurwitz empirical risk minimiser. For sampling without replacement, (5) coincides with the also well-known Horvitz-Thompson estimator of $\ell_0(\boldsymbol{\theta})$ (Horvitz and Thompson, 1952). We also note that $\hat{\ell}_{\boldsymbol{\mu}}(\boldsymbol{\theta})$ is an unbiased estimator of $\ell_0(\boldsymbol{\theta})$, provided that $\mu_i > 0$ for all $i \in \mathcal{D}$, and $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ a consistent estimator of the full-data parameter $\boldsymbol{\theta}_0$ under general regularity conditions (Binder, 1983).

An important question to ask is how the subset $\mathcal{S}$ used for the approximate solution (4) to the problem (1)–(2) should be selected for optimal performance. The problem of optimal subsampling has a long standing tradition within the field of survey sampling for inference regarding finite populations; see, e.g., Neyman (1938); Hájek (1959); Cassel et al. (1976); Brewer (1979) and Bellhouse (1984). Their work, however, is primarily concerned with linear estimators of scalar finite population characteristics. Stimulated by modern technological developments, the question of optimal subdata selection has attained renewed attention during the past few years also for more complex inference problems, as outlined above. 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 (Imberg et al., 2020; 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.

We present a theory of optimal design for general data subsampling problems, including finite population inference, parametric density estimation, and regression modelling using quasi-likelihood methods. We derive optimality conditions for a broad class of optimality criteria, including A-, D-, E-, L-, and Kiefer's Φ_q-optimality criterion (Kiefer, 1974). Algorithms to find optimal sampling schemes are presented for Poisson sampling and multinomial sampling designs. 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 affine transformations of the data and re-parameterisation of the model. The presented methodology and algorithms are illustrated in an application in the traffic safety domain.

We start with a brief review of some standard methods in unequal probability sampling and optimal design in Section 2. A general theory of optimal design for data subsampling problems is presented in Section 3, including algorithms for finding optimal sampling schemes. We discuss optimal design from a distance-minimising perspective in Section 4, and present optimal designs for some common statistical distance functions. Comments on the implementation of optimal subsampling methods in practice are provided in Section 5. Examples and experiments are presented in Section 6. We refer to Appendix A for proofs.

2 Preliminaries

Consider a class of experiments Ξ and corresponding consistent estimators $\hat{\boldsymbol{\theta}}_\xi, \xi \in \Xi$, for an unknown parameter $\boldsymbol{\theta}^*$. The aim of optimal design is to find an experiment $\xi \in \Xi$ that minimises some suitable function Φ of the covariance matrix of the estimator $\hat{\boldsymbol{\theta}}_\xi$. For instance, Φ may be the sum or product of the eigenvalues of its matrix argument, corresponding to A- or D-optimality (Atkinson and Donev, 1992), or some other measure of "size" of a matrix.

In the context of data subsampling, the experiment is determined by the choice of sampling design and sampling scheme $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$. For the estimation problem outlined in Section 1, we wish to find a sampling scheme $\boldsymbol{\mu}$ that minimises $\Phi(\mathbf{Cov}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}))$ for some suitable family of sampling designs and objective function $\Phi : \mathbb{R}^{p \times p} \to \mathbb{R}$. Some common unequal probability sampling designs are presented in Section 2.1. Expressions for the approximate covariance matrix of the estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ are provided in Section 2.2, and a brief review of optimal design in Section 2.3.

2.1 Unequal probability sampling designs

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. Following the notation in Section 1, we let S_i be the number of times an element $i \in \mathcal{D}$ is selected by the sampling mechanism, where sampling may be with or without replacement, and μ_i be the corresponding expected number of selections. We let n denote the expected size of the subsample, and $\mathcal{M}_n$ the corresponding domain of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$, i.e., the set of feasible values of the sampling scheme $\boldsymbol{\mu}$ within a specified family of sampling designs of (expected) size n. We assume that sampling is conducted according to one of the following families of sampling designs:

i) Poisson sampling with replacement (PO-WR): $S_1, \ldots, S_N$ are independent with $S_i \sim \text{Poisson}(\mu_i)$, $\mu_i > 0$. The sample size $\sum_{i \in \mathcal{D}} S_i$ is random, with expectation $\text{E}[\sum_{i \in \mathcal{D}} S_i] = \sum_{i \in \mathcal{D}} \mu_i = n$. The corresponding domain $\mathcal{M}_n$of $\boldsymbol{\mu}$ is given by $\mathcal{M}_n = \{\boldsymbol{\mu} \in \mathbb{R}^N : \mu_i > 0 \text{ for all } i \in \mathcal{D} \text{ and } \sum_{i \in \mathcal{D}} \mu_i = n\}$.

ii) Poisson sampling without replacement (PO-WOR): $S_1, \ldots, S_N$ are independent with $S_i \sim \text{Bernoulli}(\mu_i)$, $\mu_i \in (0, 1]$. The sample size $\sum_{i \in \mathcal{D}} S_i$ is random, with expectation $\text{E}[\sum_{i \in \mathcal{D}} S_i] = \sum_{i \in \mathcal{D}} \mu_i = n$. The corresponding domain $\mathcal{M}_n$of $\boldsymbol{\mu}$ is given by $\mathcal{M}_n = \{\boldsymbol{\mu} \in \mathbb{R}^N : \mu_i \in (0, 1] \text{ for all } i \in \mathcal{D} \text{ and } \sum_{i \in \mathcal{D}} \mu_i = n\}$.

iii) Multinomial sampling (MULTI): $(S_1, \ldots, S_N) \sim \text{Multinomial}(n, \boldsymbol{\mu}/n)$, $n \in \mathbb{N}, \mu_i > 0, \sum_{i \in \mathcal{D}} \mu_i = n$. Sampling is done with replacement and the sample size is fixed, i.e., $\sum_{i \in \mathcal{D}} S_i = n$. The corresponding domain $\mathcal{M}_n$ of $\boldsymbol{\mu}$ is given by $\mathcal{M}_n = \{\boldsymbol{\mu} \in \mathbb{R}^N : \mu_i > 0 \text{ for all } i \in \mathcal{D} \text{ and } \sum_{i \in \mathcal{D}} \mu_i = n\}$.

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 without replacement and with fixed selection probabilities, for instance using conditional Poisson sampling (Hájek, 1981; Tillé, 2006). This method is, however, both computationally and analytically intractable, and will therefore not be considered in this paper. Additional details may be found in, e.g., Tillé (2006) and Fuller (2009).

2.2 Covariance matrix of the Hansen-Hurwitz empirical risk minimiser

Binder (1983) showed that under suitable regularity conditions the distribution of the estimator (4) with respect to the sampling mechanism is approximately Gaussian with mean

$$\mathrm{E}[\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}] = \boldsymbol{\theta}_0 + o(n^{-1/2}), \tag{6}$$

and covariance matrix

$$\mathbf{Cov}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}} - \boldsymbol{\theta}_0) = \boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0) + o(n^{-1}), \quad \boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0) = \mathbf{H}(\boldsymbol{\theta}_0)^{-1}\mathbf{V}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)\mathbf{H}(\boldsymbol{\theta}_0)^{-1}. \tag{7}$$

Here $o(n^{-1/2})$ and $o(n^{-1})$ are interpreted elementwise and $\mathbf{H}(\boldsymbol{\theta}_0) = \frac{\partial^2 \ell_0(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\mathsf{T}}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ is the Hessian of the full-data empirical risk function (2) at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

$$\mathbf{V}(\boldsymbol{\mu}; \boldsymbol{\theta}_0) = \mathbf{Cov}\left(\nabla_{\boldsymbol{\theta}} \hat{\ell}_{\boldsymbol{\mu}}(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right) = \sum_{i,j \in \mathcal{D}} \frac{\mathrm{Cov}(S_i, S_j)}{\mu_i \mu_j} \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_j(\boldsymbol{\theta}_0)^\mathsf{T} \tag{8}$$

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

It follows from the properties of the sampling designs described in Section 2.1, that the matrix $\mathbf{V}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$ can be simplified to

$$\mathbf{V}(\boldsymbol{\mu}; \boldsymbol{\theta}_0) = \begin{cases} \sum_{i \in \mathcal{D}} \mu_i^{-1} \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}, & \text{for PO-WR and MULTI designs, and} \\ \sum_{i \in \mathcal{D}} (\mu_i^{-1} - 1) \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}, & \text{for PO-WOR.} \end{cases} \tag{9}$$

See, e.g., Tillé (2006). To obtain the above result for the multinomial sampling design, we have also used (3).

2.3 Optimal design

For an unknown parameter $\boldsymbol{\theta}^*$, consider a class of experiments Ξ and corresponding consistent estimators $\hat{\boldsymbol{\theta}}_\xi, \xi \in \Xi$, with unequal covariance matrices $\boldsymbol{\Gamma}_\xi$. 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

$$\Phi(\mathbf{U}) \geq \Phi(\mathbf{V}) \text{ for all } \mathbf{U}, \mathbf{V} \in \boldsymbol{S}_+^{p \times p} \text{ such that } \mathbf{U} \geq \mathbf{V}, \tag{10}$$

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

Some popular optimality criteria are defined and summarised in Table 1. These include the D-optimality criterion (minimise the determinant of the covariance matrix), the E-optimality criterion (minimise the largest eigenvalue of the covariance matrix) and the 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 $\mathbf{L} = \mathbf{c}$ for some $p \times 1$ vector $\mathbf{c}$, respectively (Silvey, 1980; Atkinson and Donev, 1992). Included in Table 1 is also the Φ_q- and $\Phi_{q,\mathbf{A}}$-optimality criteria, which encompass all other optimality criteria in this table. In particular, Φ_q-optimality coincides with D-optimality when $q = 0$, A-optimality when $q = 1$, and E-optimality when $q = \infty$ (Kiefer, 1974). Hence, Φ_q-optimality can be used to interpolate between A-, D- and E-optimality.

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 (parallelepiped) around the confidence ellipsoid (Pronzato and Pázman, 2013).

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). 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 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 for 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 parameterisation- and transformation-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 examples of transformation- and parameterisation-invariant linear optimality criteria will be discussed in Section 4.3.

For further details, were refer to Silvey (1980), Atkinson and Donev (1992) and Pukelsheim (1993) and Pronzato and Pázman (2013).

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

Optimality criterion	**Description**	**Objective function** $\Phi(\mathbf{\Gamma})$
A-optimality	Minimise average variance, minimise trace of covariance matrix, minimise sum of eigenvalues.	$\frac{1}{p}\sum_{i=1}^{p} \mathrm{Var}(\hat{\theta}_i) = \frac{1}{p}\mathrm{tr}(\mathbf{\Gamma})$
c-optimality	Minimise variance of a linear combination or contrast $\mathbf{c}^\mathsf{T}\hat{\boldsymbol{\theta}}$.	$\mathrm{Var}(\mathbf{c}^\mathsf{T}\hat{\boldsymbol{\theta}}) = \mathbf{c}^\mathsf{T}\mathbf{\Gamma}\mathbf{c} = \mathrm{tr}(\mathbf{\Gamma}\mathbf{c}\mathbf{c}^\mathsf{T})$
D-optimality	Minimise generalised variance, minimise determinant of covariance matrix, minimise product of eigenvalues.	$\det(\mathbf{\Gamma})^{1/p}$ or $\log\det(\mathbf{\Gamma})$
$\mathrm{D_A}$-optimality	Minimise generalised variance for subset of parameters, collection of linear combinations, or contrasts $\mathbf{A}^\mathsf{T}\boldsymbol{\theta}$.	$\det(\mathbf{A}^\mathsf{T}\mathbf{\Gamma}\mathbf{A})$
E-optimality	Minimise maximal eigenvalue, minimise variance along the direction of largest uncertainty.	$\lambda_{\max}(\mathbf{\Gamma})$
L-optimality	Minimise average variance of a collection of linear combinations or contrasts $\mathbf{L}^\mathsf{T}\boldsymbol{\theta}$. $\mathbf{L} = \boldsymbol{c} \quad \Leftrightarrow$ c-optimality $\mathbf{L} = \mathbf{I}_{p \times p} \Leftrightarrow$ A-optimality	$\frac{1}{m}\sum_{i=1}^{m} \mathrm{Var}(\mathbf{a}_i^\mathsf{T}\hat{\boldsymbol{\theta}}) = \frac{1}{m}\mathrm{tr}(\mathbf{\Gamma}\mathbf{L}\mathbf{L}^\mathsf{T})$
V-optimality	Minimise average prediction variance with respect to a measure $d\nu(\boldsymbol{x})$ on $\mathcal{X}$, assuming a linear model $\hat{y} = \boldsymbol{\varphi}(\boldsymbol{x})^\mathsf{T}\hat{\boldsymbol{\theta}}$.	$\int_{\mathcal{X}} \mathrm{Var}(\boldsymbol{\varphi}(\boldsymbol{x})^\mathsf{T}\hat{\boldsymbol{\theta}})d\nu(\boldsymbol{x}) =$ $\mathrm{tr}\left(\mathbf{\Gamma}\int_{\mathcal{X}} \boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x})^\mathsf{T} d\nu(\boldsymbol{x})\right)$
$\Phi_{q,\mathbf{A}}$-optimality $q \in [0, \infty]$	$\Phi_0 \quad \Leftrightarrow$ D-optimality $\Phi_{0,\mathbf{A}} \Leftrightarrow$ $\mathrm{D_A}$-optimality $\Phi_1 \quad \Leftrightarrow$ A-optimality $\Phi_{1,\mathbf{L}} \Leftrightarrow$ L-optimality $\Phi_\infty \quad \Leftrightarrow$ E-optimality	$\Phi_{q,\mathbf{A}}(\mathbf{\Gamma}) \;= \frac{1}{m}\mathrm{tr}[(\mathbf{A}^\mathsf{T}\mathbf{\Gamma}\mathbf{A})^q]^{1/q}, q \in (0, \infty)$ $\Phi_{0,\mathbf{A}}(\mathbf{\Gamma}) \;= \lim_{q \downarrow 0} \Phi_{q,\mathbf{A}}(\mathbf{\Gamma})$ $\Phi_{\infty,\mathbf{A}}(\mathbf{\Gamma}) = \lim_{q \uparrow \infty} \Phi_{q,\mathbf{A}}(\mathbf{\Gamma})$ $\Phi_q(\mathbf{\Gamma}) \quad\;\, = \Phi_{q,\mathbf{A}}(\mathbf{\Gamma}), \mathbf{A} = \mathbf{I}_{p \times p}$

3 Optimal subsampling designs

In this section we present optimal sampling schemes for a general class of optimality criteria, under an assumption of differentiability. Φ-optimality is defined in Section 3.1, where we also present three important lemmas. Optimality criteria for Poisson and multinomial sampling designs are presented in Section 3.2, and algorithms for finding optimal sampling schemes in Section 3.3.

First we note that the approximate covariance matrix $\mathbf{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$ of the estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$, as given in (7), generally depends on the full data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i \in \mathcal{D}}$ and full-data parameter $\boldsymbol{\theta}_0$. Clearly, subsampling would not be needed if such information were available at the design stage. This is a general problem in optimal design, however, and not specific to our setup, and hence not a major limitation of the theory we present. We will proceed in this section and Section 4 as if such information is available, keeping in mind that the resulting theoretically optimal designs can generally not be found in practice. We refer to Section 5 for a discussion on the implementation of optimal subsampling designs in practice.

Throughout we assume regularity conditions such that (7) holds, and that $\mathbf{H}(\boldsymbol{\theta}_0)$ is of full rank. All vectors are assumed to be column vectors, unless otherwise stated. We let $||\mathbf{u}||_2^2 = \mathbf{u}^\mathsf{T}\mathbf{u}$ denote the Euclidean norm of a vector $\mathbf{u}$. Also recall that $\psi_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta})$.

3.1 Optimality criteria

By an optimal sampling scheme $\boldsymbol{\mu}^*$, we mean the following:

Definition 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* (10) *holds. Also consider a family of unequal probability sampling designs (e.g., PO-WR, PO-WOR or MULTI) indexed by the sampling scheme $\boldsymbol{\mu}$. Let the expected size $\mathrm{E}[\sum_{i\in\mathcal{D}} S_i] = n$ be fixed, and let $\mathcal{M}_n$ denote the corresponding domain of $\boldsymbol{\mu}$. We say that a sampling scheme $\boldsymbol{\mu}^*$ is Φ-optimal if*

$$\boldsymbol{\mu}^* = \arg\min_{\boldsymbol{\mu}\in\mathcal{M}_n} \Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)),$$

where $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ is the approximate covariance matrix of $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$, as given in (7)*.*

Finding a Φ-optimal sampling scheme reduces to a non-linear, possibly non-convex, restricted optimisation problem over an $(N-1)$-dimensional hyperplane in $\mathbb{R}^N$. While this problem may be addressed by numerical optimisation methods when N is small, this is generally not a viable option for large datasets. We therefore need a theory of optimal design that can be used to devise efficient algorithms for finding optimal sampling schemes when N is large. To make the problem tractable, we will restrict ourselves to optimality criteria $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ that are differentiable with respect to $\boldsymbol{\mu}$ in a neighbourhood of its optimum $\boldsymbol{\mu}^*$. Three important lemmas are provided below.

Lemma 1 (The chain rule)**.** *Consider a function $\Phi : \boldsymbol{S}_+^{p\times p} \to \mathbb{R}^p$, and assume that $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ is differentiable with respect to $\boldsymbol{\mu}$ in a neighbourhood of some point $\boldsymbol{\mu}^*$. The partial derivative of $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ with respect to μ_i is then given by*

$$\frac{\partial\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i} = \mathrm{tr}\left(\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))\frac{\partial\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)}{\partial\mu_i}\right), \tag{11}$$

where $\boldsymbol{\phi}(\mathbf{U}) = \frac{\partial\Phi(\mathbf{U})}{\partial\mathbf{U}}$ is the $p\times p$ matrix derivative of Φ with respect to its matrix argument, and $\frac{\partial\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)}{\partial\mu_i}$ is the elementwise derivative of $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ with respect to μ_i.

Assume further that

- *i) $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ decreases monotonically with $\mu_1,\ldots,\mu_N$ in the Loewner order sense, i.e., $\boldsymbol{\Gamma}(\boldsymbol{\mu}_1;\boldsymbol{\theta}_0) - \boldsymbol{\Gamma}(\boldsymbol{\mu}_2;\boldsymbol{\theta}_0)$ is positive semi-definite for every pair of vectors $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2 \in \mathbb{R}_{>0}^N$ such that $\boldsymbol{\mu}_1 \leq \boldsymbol{\mu}_2$ (elementwise), and*
- *ii) Φ is monotone for Loewner's ordering, i.e., that* (10) *holds.*

Then the matrix $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ is positive semi-definite and there exists a real matrix $\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ such that $\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^\mathsf{T} = \boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$.

The first part of Lemma 1 follows by the chain rule in matrix differential calculus and the symmetry of $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$, and the second by the monotonicity assumptions on $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ and Φ. The matrix $\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ may, e.g., be obtained as the matrix square root of $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$, or by the Cholesky decomposition when $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ is of full rank. Some examples are provided in Lemma 2.

Lemma 2 ($\boldsymbol{\mu}$-differentiable Φ-optimality criteria)**.** *Consider a PO-WR, PO-WOR or MULTI design, and assume that $\mathbf{H}(\boldsymbol{\theta}_0)$ is of full rank. Let $\mathbf{c}$ be a non-zero $p\times 1$ vector, $\mathbf{L}$ a non-zero $p\times m$ matrix, $\lambda_{\max}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ the maximal eigenvalue of $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$, and $\boldsymbol{v}_{\boldsymbol{\mu}}$ a corresponding eigenvector. Let $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ be defined as in Lemma 1. Then the following holds:*

- *a) $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ is differentiable with respect to $\boldsymbol{\mu}$ and $\frac{\partial\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)}{\partial\mu_i} = -\mu_i^{-2}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}$, provided that $\mu_i > 0$.*
- *b) The D-optimality objective function $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \log\det(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ is differentiable with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{-1}$, provided that $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ is of full rank.*
- *c) The E-optimality objective function $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \lambda_{\max}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ is differentiable with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \boldsymbol{v}_{\boldsymbol{\mu}}\boldsymbol{v}_{\boldsymbol{\mu}}^\mathsf{T}$, provided that $\lambda_{\max}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ has multiplicity* 1*.*

d) *The L-optimality objective function* $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \operatorname{tr}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{L}\mathbf{L}^{\mathsf{T}})$ *is differentiable with respect to* $\boldsymbol{\mu}$*, and* $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \mathbf{L}\mathbf{L}^{\mathsf{T}}$*. In particular, this holds for A-optimality with* $\mathbf{L} = \mathbf{I}_{p\times p}$ *and c-optimality with* $\mathbf{L} = \mathbf{c}$.

e) *The* Φ_q*-optimality objective function* $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \operatorname{tr}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^q)^{1/q}$ *is differentiable with respect to* $\boldsymbol{\mu}$ *for* $q \in (0,\infty)$ *and* $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \operatorname{tr}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^q)^{1/q-1}\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{q-1}$*, provided that* $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ *is of full rank.*

Combining the results of Lemma 1 and 2, we obtain the following:

Lemma 3 (Partial derivatives of $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$)**.** *Consider a PO-WR, PO-WOR or MULTI design. Also consider a function* $\Phi : \boldsymbol{S}_+^{p\times p} \to \mathbb{R}$ *such that* Φ *is monotone for Loewner's ordering. Assume that* $\mathbf{H}(\boldsymbol{\theta}_0)$ *is of full rank, and that* $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ *is differentiable with respect to* $\boldsymbol{\mu}$ *in a neighbourhood of some point* $\boldsymbol{\mu}^*$*. Let* $\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ *be defined as in Lemma 1. Then*

$$\frac{\partial\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i} = -\mu_i^{-2}\big|\big|\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{\mathsf{T}}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\big|\big|_2^2.$$

3.2 Optimality conditions

Using results of Lemma 1–3, in Proposition 1 we present optimality conditions for Poisson and multinomial sampling designs with respect to a Φ-optimality criterion under an assumption of differentiability.

Proposition 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. Assume that* $\mathbf{H}(\boldsymbol{\theta}_0)$ *is of full rank, and that* $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ *is differentiable with respect to* $\boldsymbol{\mu}$ *in a neighbourhood of some point* $\boldsymbol{\mu}^*$*. Let* $\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ *be defined according to Lemma 1, and*

$$c_i = \big|\big|\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)^{\mathsf{T}}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\big|\big|_2^2. \tag{12}$$

Then the following holds:

a) $\boldsymbol{\mu}^*$ *is a stationary point of* $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ *for a PO-WR or MULTI design of size* n *if*

$$\mu_i^* = n\frac{\sqrt{c_i}}{\sum_{j\in\mathcal{D}}\sqrt{c_j}} \quad \textit{for all } i \in \mathcal{D}. \tag{13}$$

b) $\boldsymbol{\mu}^*$ *is a stationary point of* $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ *for a PO-WOR design of size* n *if*

$$\mu_i^* \leq 1 \qquad \textit{for all } i \in \mathcal{D}, \tag{14a}$$

$$\mu_i^* = (n - n_{\mathcal{E}})\frac{\sqrt{c_i}}{\sum_{j\in\mathcal{D}\setminus\mathcal{E}}\sqrt{c_j}} \qquad \textit{for all } i \in \mathcal{D}\setminus\mathcal{E}, \tag{14b}$$

$$\sqrt{c_i} \geq \sqrt{c_j}/\mu_j^* \qquad \textit{for all } i \in \mathcal{E} \textit{ and } j \in \mathcal{D}\setminus\mathcal{E}, \tag{14c}$$

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* $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ *is convex in* $\boldsymbol{\mu}$*, then* $\boldsymbol{\mu}^*$ *is the global minimiser of* $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$.

We note that the matrix $\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)$ in Proposition 1 exists by Lemma 1 whenever the objective function is differentiable at $\boldsymbol{\mu}^*$. It need not be unique, however, and may depend on both $\boldsymbol{\mu}^*$ and $\boldsymbol{\theta}_0$. Some examples can be found in Lemma 2. For linear optimality criteria, the matrix $\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)$ does not depend on $\boldsymbol{\mu}^*$ but may depend on the full-data parameter $\boldsymbol{\theta}_0$; see Section 4.3 for further discussion and examples.

The result of Proposition 1 follows from Lemma 3 by the Lagrange multiplier method in a) and the Karush-Kuhn-Tucker conditions in b). We show in Proposition 2 that the D- and L-optimality criteria are convex in $\boldsymbol{\mu}$ and hence that global optimality can be deduced.

Proposition 2 (Convexity of the D- and L-optimality criteria)**.** *Consider the family of PO-WR, PO-WOR or multinomial sampling designs of (expected) size* n*. Assume that* $\mathbf{H}(\boldsymbol{\theta}_0)$ *is of full rank. Then*

a) *the L-optimality criterion is convex in* $\boldsymbol{\mu}$.

Assume further that $\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$*, defined in* (8)*, is positive definite for every* $\boldsymbol{\mu} \in \mathcal{M}_n$*. Then* *b) the D-optimality criterion is (log) convex in $\boldsymbol{\mu}$.*

The first assumption in Proposition 2 is needed to ensure that the inverse of $\mathbf{H}(\boldsymbol{\theta}_0)$ exists, and that the approximate covariance matrix $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ is well-defined. For the D-optimality criterion we also need that $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ is of full rank, which follows if the additional assumption on $\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ is fulfilled. We note that this is rather an assumption on the model and data than on the sampling design. Moreover, both of the assumptions in Proposition 2 hold in most situations. One example where these assumptions are violated, however, is encountered in (multivariate) regression analysis when the model matrix $\mathbf{X}$ or response matrix $\mathbf{Y}$ (i.e., the matrices with rows $\boldsymbol{x}_i^\mathsf{T}$ and $\boldsymbol{y}_i^\mathsf{T}$) has linearly dependent columns. Another example is logistic regression with complete separation, i.e., when the outcome is linearly separable by the predictors. It is also possible that $\mathbf{H}(\boldsymbol{\theta}_0)$ is of full rank while $\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ is rank-deficient on $\mathcal{M}_n$. In this case the L-optimality criterion is still well-defined, whereas the D-optimality criterion is not. There are various solutions to such problems, e.g, removing redundant columns from the data, using a ridge penalty to avoid rank-deficiency of the Hessian matrix (Hastie, 2020), or by restricting the D-optimality criterion to a subset of the parameters using so called D$_\mathrm{A}$-optimality (Table 1) (Sibson, 1974). Most of these situations may be avoided by a careful construction of the model, however.

Even with a convex objective function, it is possible that no feasible global optimum exist since the domain $\mathcal{M}_n$ is not closed. For the L-optimality criterion this happens if c_i in (12) is equals zero for some $i \in \mathcal{D}$. In this case the objective function does not depend on the corresponding μ_i and the partial derivative with respect to μ_i is equal to zero. The optimal choice would be to correspondingly set $\mu_i = 0$, but this is an unfeasible solution. For any choice of $\mu_i > 0$, it is always possible to improve the value of the objective function by reducing μ_i and distribute the regained probability mass optimally on the remaining elements in $\mathcal{D}$. The existence of a feasible global optimum can be ensured by imposing the additional restriction that $\mu_i \geq \mu_{\min}$ for all $i \in \mathcal{D}$, and some $\mu_{\min} \in (0, n/N)$. An alternative solution that does not require explicit specification of a lower bound $\mu_{\min}$, but that still ensures a feasible solution with $\mu_i > 0$, is proposed in Section 5.

3.3 Optimal sampling schemes

In this subsection we present algorithms for finding optimal sampling schemes. First consider a linear optimality criterion with respect to a $p \times m$ matrix $\mathbf{L}$. In this case a closed solution for the optimal sampling scheme is available for the PO-WR and MULTI designs, and given by (12)–(13) with $\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0) = \mathbf{L}$, provided that the corresponding $c_i > 0$ for all $i \in \mathcal{D}$. In particular, A-optimality is obtained with $\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0) = \mathbf{I}_{p\times p}$, and c-optimality with $\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0) = \mathbf{c}$. For PO-WOR, a simple adjustment may be needed to ensure that a feasible solution with $\mu_i \leq 1$ is obtained (Algorithm 1).

Algorithm 1. L-optimal sampling schemes for Poisson and multinomial sampling designs.

INPUT: Index set $\mathcal{D}$, (expected) sample size n, non-zero $p \times m$ matrix $\mathbf{L}$, Hessian matrix $\mathbf{H}(\boldsymbol{\theta}_0)$, gradients $\{\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\}_{i\in\mathcal{D}}$, family of sampling designs (PO-WR, PO-WOR or MULTI).

```
Let $c_i = ||\mathbf{L}^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)||_2^2$ 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
```

Using the result of Proposition 1 and Algorithm 1, in Algorithm 2 we present an iterative algorithm 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. If the algorithm converges, it converges to a fixed-point of the function $\boldsymbol{h}(\boldsymbol{u}) : \mathbb{R}^N \to \mathbb{R}^N$ defined by Algorithm 1 with $\mathbf{L} = \mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$, which by Proposition 1 is a stationary point of $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$. 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. The performance of this algorithm for non-linear optimality criteria will be evaluated in Section 6.

Algorithm 2. Fixed-point iteration.

```
INPUT: Index set $\mathcal{D}$, (expected) sample size $n$, optimality criterion $\Phi$, Hessian matrix $\mathbf{H}(\boldsymbol{\theta}_0)$, gradients $\{\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\}_{i\in\mathcal{D}}$, 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 $\mathbf{L}_t\mathbf{L}_t^\top = \boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu}_{t-1};\boldsymbol{\theta}_0))$.
    Let $c_i = ||\mathbf{L}_t^\top\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)||_2^2$ 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 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
```

4 A distance-minimising perspective on optimal subsampling designs

Recall the overall aim of data subsampling as introduced in Section 1; to find an approximate solution to the originally intractable problem (1)–(2). A natural target for optimal design in this context is therefore to minimise the expected distance $\mathrm{E}[d(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}})]$ of the estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ from the full-data parameter $\boldsymbol{\theta}_0$, for some suitable statistical distance function $d : \boldsymbol{\Omega} \to \mathbb{R}_+$. In Section 4.1 we define a class of optimality criteria for minimising the expected distance, and discuss their relation to traditional optimality criteria. Some specific examples are presented in Section 4.2, and invariance properties discussed in Section 4.3.

4.1 *d*-optimality

Consider a statistical distance function $d(\boldsymbol{\theta})$ such that $d(\boldsymbol{\theta}) \geq 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Omega}$, with equality only for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. For analytical and computational tractability we also require the distance function to be twice differentiable, and let $\mathbf{H}_d(\boldsymbol{\theta}) = \frac{\partial^2 d(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}$ denote the Hessian matrix of $d(\boldsymbol{\theta})$. We have the following result:

Lemma 4 (Taylor expansion of $d(\boldsymbol{\theta})$)**.** *Let $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ be defined according to* (1)–(2) *and* (4)–(5)*. Assume that* (6)–(7) *hold, and that $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ has bounded $2+\delta$ moments for some $\delta > 0$. Consider a function $d : \boldsymbol{\Omega} \to \mathbb{R}_+$ such that $d(\boldsymbol{\theta}) = 0$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Assume that $d(\boldsymbol{\theta})$ is twice differentiable in a neighbourhood of $\boldsymbol{\theta}_0$, and that $\mathbf{H}_d(\boldsymbol{\theta}_0)$ is non-zero. Then*

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

The result of Lemma 4 follows from a Taylor expansion of $d(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and properties of quadratic forms. Based on this result, we define a class of expected-distance-minimising optimality criteria as follows:

Definition 2 (*d*-optimality)**.** *Consider a function $d : \boldsymbol{\Omega} \to \mathbb{R}_+$ satisfying the conditions of Lemma 4. Also consider a family of unequal probability sampling designs (e.g., PO-WR, PO-WOR or MULTI) indexed by the sampling scheme $\boldsymbol{\mu}$. Let the expected size $\mathrm{E}[\sum_{i\in\mathcal{D}} S_i] = n$ be fixed, and let $\mathcal{M}_n$ denote the corresponding domain of $\boldsymbol{\mu}$. We say that a sampling scheme $\boldsymbol{\mu}^*$ is d-optimal with respect to the statistical distance function $d(\boldsymbol{\theta})$ if*

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

We denote this optimality criterion as *d*-optimality for *distance*, which should not be confused with the D-optimality criterion introduced in Section 2.3. We recognise the *d*-optimality criterion as a linear optimality criterion with $\mathbf{L}\mathbf{L}^\mathsf{T} = \mathbf{H}_d(\boldsymbol{\theta}_0)$. Indeed, we have the following equivalence result:

Proposition 3 (Equivalence between d- and Φ-optimality)**.**

a) Consider a function $d : \boldsymbol{\Omega} \to \mathbb{R}_+$ satisfying the conditions of Lemma 4 and denote by $\mathbf{H}_d(\boldsymbol{\theta})$ the Hessian of $d(\boldsymbol{\theta})$. Assume that the sampling scheme $\boldsymbol{\mu}^*$ is d-optimal with respect to the distance function $d(\boldsymbol{\theta})$. Then there exists a real matrix $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}^\mathsf{T} = \mathbf{H}_d(\boldsymbol{\theta}_0)$ and $\boldsymbol{\mu}^*$ is L-optimal with respect to $\mathbf{L}$.

b) Let $\Phi : \boldsymbol{S}_+^{p\times p} \to \mathbb{R}$ and $\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ be defined as in Lemma 1 and assume that $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ is differentiable with respect to $\boldsymbol{\mu}$ in a neighbourhood of its optimum argument $\boldsymbol{\mu}^*$. Then $\boldsymbol{\mu}^*$ is d-optimal with respect to the distance function $d(\boldsymbol{\theta}) = ||\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)^\mathsf{T}(\boldsymbol{\theta} - \boldsymbol{\theta}_\mathbf{0})||_2^2$.

Proposition 3 follows immediately by the definitions and the optimality conditions of Proposition 1. By this result, any Φ-optimality criterion may be viewed as minimising the expected distance of the estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ from the full-data parameter $\boldsymbol{\theta}_0$ for a particular choice of distance function. For instance, A-optimality is equivalent to *d*-optimality with $d(\boldsymbol{\theta}) = ||\boldsymbol{\theta}-\boldsymbol{\theta}_0||_2^2$. Beyond linear optimality criteria, the induced distance function may be implicit and depend on the Φ-optimal sampling scheme $\boldsymbol{\mu}^*$. As an example, E-optimality is equivalent to *d*-optimality with $d(\boldsymbol{\theta}) = ||\boldsymbol{v}_{\boldsymbol{\mu}^*}^\mathsf{T}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)||_2^2$, where $\boldsymbol{v}_{\boldsymbol{\mu}^*}$ is an eigenvector pertaining to the largest eigenvalue of $\boldsymbol{\Gamma}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)$ 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.

4.2 Some distance-minimising designs

Next we show how *d*-optimality may be used to derive a novel class of linear optimality criteria with good theoretical properties, including transformation- and parameterisation invariance. Consider the following statistical distance functions naturally arising in data subsampling applications and commonly encountered in statistics:

i) **Empirical risk distance**: Since $\boldsymbol{\theta}_0$ is defined as the minimiser of the full-data empirical risk (2), we may measure of the distance of a parameter value $\boldsymbol{\theta}$ from the full-data parameter $\boldsymbol{\theta}_0$ through the attained value of the empirical risk. We define the empirical risk distance of $\boldsymbol{\theta}$ from $\boldsymbol{\theta}_0$ as $d_{\mathrm{ER}}(\boldsymbol{\theta}) = \ell_0(\boldsymbol{\theta}) - \ell_0(\boldsymbol{\theta}_0)$.

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

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 $\boldsymbol{\theta} \in \mathbb{R}^p$ from the mean $\boldsymbol{\gamma}$ is then given by $\sqrt{(\boldsymbol{\theta}-\boldsymbol{\gamma})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\gamma})}$. We define the squared Mahalanobis distance of $\boldsymbol{\theta}$ from $\boldsymbol{\theta}_0$ with respect to a real, symmetric, positive definite dispersion matrix $\boldsymbol{\Sigma}$ as $d_{\boldsymbol{\Sigma}}(\boldsymbol{\theta}) = (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$.

Four natural choices of the dispersion matrix $\boldsymbol{\Sigma}$ for the Mahalanobis distance are:

iii.a) $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$, the approximate covariance matrix of $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$.

iii.b) $\boldsymbol{\Sigma} = \mathbf{H}(\boldsymbol{\theta}_0)^{-1}$, which for a parametric model is an estimate of the covariance matrix of $\boldsymbol{\theta}_0$, seen as an estimator of some underlying super-population parameter $\boldsymbol{\theta}^*$. In this case, $\mathbf{H}(\boldsymbol{\theta}_0)$ is also known as the observed Fisher information matrix, often denoted as $\mathbf{I}(\boldsymbol{\theta}_0)$ (Efron and Hinkley, 1978).

iii.c) $\boldsymbol{\Sigma} = \widetilde{\mathbf{H}}(\boldsymbol{\theta}_0)^{-1}$, where $\widetilde{\mathbf{H}}(\boldsymbol{\theta}_0)$ is defined for a parametric model $f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$ as $\widetilde{\mathbf{H}}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{y}\sim f_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x})}[\mathbf{H}(\boldsymbol{\theta}_0)]$. In this case, $\widetilde{\mathbf{H}}(\boldsymbol{\theta}_0)$ is also known as the expected Fisher information matrix, often denoted as $\mathcal{I}(\boldsymbol{\theta}_0)$ (Efron and Hinkley, 1978).

iii.d) $\boldsymbol{\Sigma} = \mathbf{H}(\boldsymbol{\theta}_0)^{-1}\mathbf{V}(\boldsymbol{\theta}_0)\mathbf{H}(\boldsymbol{\theta}_0)^{-1}$, with

$$\mathbf{V}(\boldsymbol{\theta}_0) = \sum_{i\in\mathcal{D}} \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}, \quad \boldsymbol{\psi}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}). \tag{15}$$

This choice of the matrix $\boldsymbol{\Sigma}$ corresponds to the "robust estimator" or "sandwich estimator" of the covariance matrix of $\boldsymbol{\theta}_0$, seen as an estimator of some underlying super-population parameter $\boldsymbol{\theta}^*$ under a semi-parametric or presumably misspecified parametric model (Stefanski and Boos, 2002).

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. We also define d_{I}-, $d_{\mathcal{I}}$- and d_{S}-optimality as $d_{\boldsymbol{\Sigma}}$-optimality with dispersion matrix $\boldsymbol{\Sigma}$ taken as in iii.b) (the inverse of the observed information matrix), iii.c) (the inverse of the expected information matrix) and iii.d) (the sandwich variance estimator), respectively.

Note that d_{KL}- and $d_{\mathcal{I}}$-optimality are defined for parametric models only, whereas d_{ER}-, d_{I}- and d_{S}-optimality are appropriate also for semi-parametric and distribution-free methods, including estimation of finite population characteristics. For regression problems, the *d*-optimality criterion with the empirical risk distance (i.e., d_{ER}-optimality) is closely related to the V-optimality criterion (Table 1, Section 2.3). Indeed, these two optimality criteria are equivalent for ordinary least squares regression when $\nu(\boldsymbol{x})$ is the empirical measure on $\{\boldsymbol{x}_i\}_{i\in\mathcal{D}}$.

The Mahalanobis distance with $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$ arises by considering the uncertainty of $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ as an estimator of the full-data parameter $\boldsymbol{\theta}_0$. In contrast, our motivation for the dispersion matrices in iii.b)–iii.d) above comes from a super-population viewpoint where $\boldsymbol{\theta}_0$ is seen as an estimator of some underlying parameter $\boldsymbol{\theta}^*$ (cf. Hartley and Sielken, 1975). The different choices of dispersion matrix $\boldsymbol{\Sigma}$ then arise naturally trough different measures of uncertainty associated with the full-data parameter $\boldsymbol{\theta}_0$ (cf. Stefanski and Boos, 2002). We emphasise, however, that the super-population perspective adopted here is purely rhetorical. The resulting distance functions are equally valid even without any intentions of super-population inference. The significance of these particular choices of distance functions and dispersion matrices are highlighted in Proposition 4 below and further in Section 4.3.

Proposition 4 (d_{ER}- d_{KL}-, $d_{\boldsymbol{\Sigma}}$-optimality and equivalence with L-optimality)**.**

a) d-optimality with respect to the empirical risk distance is equivalent to L-optimality with respect to a $p \times p$ *matrix* $\mathbf{L}$ *such that* $\mathbf{L}\mathbf{L}^{\mathsf{T}} = \mathbf{H}(\boldsymbol{\theta}_0)$.

b) d-optimality with respect to the Mahalanobis distance is equivalent to L-optimality with respect to a $p \times p$ *matrix* $\mathbf{L}$ *such that* $\mathbf{L}\mathbf{L}^{\mathsf{T}} = \boldsymbol{\Sigma}^{-1}$.

c) Consider a parametric statistical model with density function $f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$ *and cumulative distribution function* $F_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$*. Let* $\boldsymbol{\theta}_0$ *be defined by* (1)–(2) *with* $\ell_i(\boldsymbol{\theta}) = -\log f_{\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{x}_i)$*. Assume that the following holds for all* $i \in \mathcal{D}$ *and all parameter values* $\boldsymbol{\theta}$ *in a neighbourhood or* $\boldsymbol{\theta}_0$*:* $d_{\mathrm{KL}}(\boldsymbol{\theta})$ *is finite,* $\ell_i(\boldsymbol{\theta})$ *is two times continuously differentiable with respect to* $\boldsymbol{\theta}$*, and all first- and second-order derivatives of* $\log f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}_i)$ *are bounded in* L_1 *with respect to the measure* $dF_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}_i)$*. Then d-optimality with respect to the Kullback-Leibler distance is equivalent to L-optimality with respect to a* $p \times p$ *matrix* $\mathbf{L}$ *such that* $\mathbf{L}\mathbf{L}^{\mathsf{T}} = \widetilde{\mathbf{H}}(\boldsymbol{\theta}_0)$.

The result of Proposition 4 follows immediately from Proposition 3. Note that for c) we need conditions on the model $f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$ that allow us to change the order of integration and differentiation.

By Proposition 4a) and c) we observe that d_{ER}- and d_{I}-optimality are equivalent (take $\boldsymbol{\Sigma} = \mathbf{H}(\boldsymbol{\theta}_0)^{-1}$). The same also holds for d_{KL}- and $d_{\mathcal{I}}$-optimality (take $\boldsymbol{\Sigma} = \widetilde{\mathbf{H}}(\boldsymbol{\theta}_0)^{-1}$). We also note that for many models, including exponential families and generalised linear models with a canonical link function, the observed information matrix $\mathbf{I}(\boldsymbol{\theta}_0) = \mathbf{H}(\boldsymbol{\theta}_0)$ and expected information matrix $\mathcal{I}(\boldsymbol{\theta}_0) = \widetilde{\mathbf{H}}(\boldsymbol{\theta}_0)$ are equal, and that these four optimality criteria hence are equivalent (see, e.g. McCullagh and Nelder, 1989). For a correctly specified parametric model, they are also asymptotically equivalent to d_{S}-optimality (as $N \to \infty$), since in this case $N\mathbf{H}(\boldsymbol{\theta}_0)^{-1}$, $N\widetilde{\mathbf{H}}(\boldsymbol{\theta}_0)^{-1}$ and $N\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\mathbf{V}(\boldsymbol{\theta}_0)\mathbf{H}(\boldsymbol{\theta}_0)^{-1}$ all converge to the same limit (see, e.g. Stefanski and Boos, 2002).

The above-mentioned optimality criteria are also related to A-optimality after an appropriate change of variables. Consider, e.g., a linear regression model, and assume that the model matrix $\mathbf{X}$ (i.e., the matrix with rows $\boldsymbol{x}_i^{\mathsf{T}}$) has orthogonal columns. Then the d_{ER}- and d_{KL}-optimality criteria are equivalent to A-optimality, since in this case $\mathbf{H}(\boldsymbol{\theta}_0) = \widetilde{\mathbf{H}}(\boldsymbol{\theta}_0) \propto \mathbf{X}^{\mathsf{T}}\mathbf{X} = \mathbf{I}_{p\times p}$. In the non-orthogonal case, the d_{ER}- and d_{KL}-optimality criteria depend on the parameterisation of the model and on the scaling of the data and correlations between the variables, through the Hessian $\mathbf{H}(\boldsymbol{\theta}_0)$. As a consequence, invariance under non-singular affine transformations of the data and under a re-parameterisation of the model is achieved (see Section 4.3). Geometrically, the A-optimality criterion minimises the expected Euclidean distance of the estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ from the full-data parameter $\boldsymbol{\theta}_0$ (Proposition 3, Section 4.1). The d_{ER}- and d_{KL}-optimality criteria minimise the expected distance with respect to the natural geometry of the model space.

Finally we consider the relation between *d*-optimality and D-optimality. These two criteria coincide if the distance function is taken as the squared Mahalanobis distance $d_{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$ with dispersion matrix $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}(\boldsymbol{\mu}^*; \boldsymbol{\theta}_0)$, where $\boldsymbol{\mu}^*$ is the D-optimal sampling scheme (see Proposition 3 and Proposition 4b)). In particular, D-optimality is equivalent to L-optimality with $\mathbf{L} = \mathbf{H}(\boldsymbol{\theta}_0)\mathbf{V}(\boldsymbol{\mu}^*; \boldsymbol{\theta}_0)^{-1/2}$, and with $\mathbf{V}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$ defined as in (8). This result is not very practical, however, since the coefficient matrix of the L-optimality criterion depends on the D-optimal sampling scheme $\boldsymbol{\mu}^*$. An optimality criterion closely related to D-optimality is L-optimality with $\mathbf{L} = \mathbf{H}(\boldsymbol{\theta}_0)\mathbf{V}(\boldsymbol{\theta}_0)^{-1/2}$, where $\mathbf{V}(\boldsymbol{\theta}_0)$ given by (15) does not depend on $\boldsymbol{\mu}$. By Proposition 4b), this is equivalent to d_{S}-optimality.

We point out that having the coefficient matrix $\mathbf{L}$ depending on the full-data Hessian $\mathbf{H}(\boldsymbol{\theta}_0)$ and parameter $\boldsymbol{\theta}_0$ is not restrictive, since all optimal designs anyway depend on unknown full-data characteristics. Methods to handle this issue will be addressed in Section 5.

4.3 Invariance properties

In addition to their appealing geometric and statistical interpretation, the expected-distance-minimising optimality criteria introduced in the previous section 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_{S}-optimality criteria have simple solutions for the optimal sampling schemes according to Algorithm 1. The invariance properties of these optimality criteria and their corresponding optimal sampling schemes are established below.

Consider a re-parameterisation $\boldsymbol{g} : \boldsymbol{\theta} \mapsto \boldsymbol{\eta}$, where $\boldsymbol{g}$ is a one-to-one differentiable mapping on the parameter space. Under such a transformation the full-data empirical risk minimiser $\boldsymbol{\theta}_0$ is equivariant in the sense that the minimiser of the induced empirical risk $\ell_0^*(\boldsymbol{\eta}) := \sum_{i\in\mathcal{D}} \ell_i(\boldsymbol{g}^{-1}(\boldsymbol{\eta}))$ is given by $\boldsymbol{\eta}_0 = \boldsymbol{g}(\boldsymbol{\theta}_0)$ (see, e.g., Casella and Berger, 2001). By similar arguments, the Hansen-Hurwitz empirical risk minimiser for $\boldsymbol{\eta}_0$ is given by $\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}} = \boldsymbol{g}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}})$. Evaluating the derivatives of the induced empirical risk $\ell_0^*(\boldsymbol{\eta})$, by (7) we obtain the covariance matrix of $\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}}$ as

$$\mathbf{Cov}(\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}} - \boldsymbol{\eta}_0) = \boldsymbol{\Gamma}_{\boldsymbol{g}}(\boldsymbol{\mu};\boldsymbol{\theta}_0) + o(n^{-1}), \quad \boldsymbol{\Gamma}_{\boldsymbol{g}}(\boldsymbol{\mu};\boldsymbol{\theta}_0) = \mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{\mathsf{T}}, \tag{16}$$

where $\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta})$ is the Jacobian of $\boldsymbol{g}$, i.e,. the matrix with rows $\nabla_{\boldsymbol{\theta}} g_i(\boldsymbol{\theta})^{\mathsf{T}}$. We say that an optimality criterion is invariant under a re-parameterisation $\boldsymbol{g} : \boldsymbol{\theta} \mapsto \boldsymbol{\eta}$ if the optimal sampling schemes for $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ and $\boldsymbol{\Gamma}_{\boldsymbol{g}}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ are equal. Invariance of the d_{ER}-, d_{KL}- and d_{S}-optimality criteria is established in Proposition 5.

Proposition 5 (Parameterisation invariance)**.** *Let $\mathbf{V}(\boldsymbol{\theta}_0)$ be defined as in* (15)*, and assume that $\mathbf{H}(\boldsymbol{\theta}_0)$ and $\mathbf{V}(\boldsymbol{\theta}_0)$ are of full rank. Then the d_{ER}- and d_{S}-optimality criteria are invariant under a re-parameterisation $\boldsymbol{g} : \boldsymbol{\theta} \mapsto \boldsymbol{\eta}$, where $\boldsymbol{g}$ is a one-to-one differentiable mapping on the parameter space. Under the assumptions of Proposition 4c), the same also holds for the d_{KL}-optimality criterion.*

Similar results may also be obtained for invariance under non-singular affine transformations of the data. Indeed, in many cases a transformation of the data induces a transformation on the parameter space that satisfies the conditions on the transformation $\boldsymbol{g}$ in Proposition 5. Care needs to be taken, however, to make sure that the empirical risk function is still defined after applying the transformation, and that the transformation produces a mathematically equivalent model. Under such circumstances, the notions of transformation- and parameterisation-invariance are interchangeable in most practical situations. Exceptions exist, however, where a transformation of the data renders the Hessian $\mathbf{H}(\boldsymbol{\theta}_0)$ unchanged. In such a case, the d_{ER}- and d_{KL}-optimality criteria are no longer invariant under affine transformations of the data. We provide such an example in Section 6.4. We note that even in such cases the D- and d_{S}-optimality criteria remain invariant under affine transformations of the data.

5 Practical implementation

Thus far, we have assumed the full data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i\in\mathcal{D}}$ and full-data parameter $\boldsymbol{\theta}_0$ to be known. However, if such information were available at the design stage, subsampling would not be needed in the first place. In this section we describe a practical approach to optimal subsampling. In Section 5.1 we introduce the anticipated covariance matrix (cf. Isaki and Fuller, 1982) to be used in the optimisation as a surrogate for the unknown covariance matrix $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$. Sequential optimal design and multi-stage sampling procedures, where the information needed for the optimisation is acquired gradually during the sampling process, are discussed in Section 5.2.

5.1 Auxiliary-variable-assisted subsampling designs

In addition to the data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i\in\mathcal{D}}$, we now assume the existence of a collection of auxiliary variables $\{\boldsymbol{z}_i\}_{i\in\mathcal{D}}$, which are available *a priori* for all members $i \in \mathcal{D}$. Depending on context, the auxiliary variables may include some of the variables in $\boldsymbol{x}_i$ and/or some of the variables in $\boldsymbol{y}_i$. For instance, consider a case-control study to investigate the effect of some exposure variables on a known binary outcome. In this case the auxiliary variables contain the (scalar) outcome y_i, and possibly some of the explanatory variables or some proxies for those (cf. Imberg et al., 2022a). The opposite situation is encountered in active learning (Settles, 2012). In this case all predictor vectors $\boldsymbol{x}_i$ are known but the outcomes y_i can be observed only for a subset $\mathcal{S} \subset \mathcal{D}$, hence $\boldsymbol{z}_i = \boldsymbol{x}_i$ (cf. Bach, 2007; Wang et al., 2017; Meng et al., 2021; Zhang et al., 2021; Imberg et al., 2022b). In the extreme case, one may even have access to the full-data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i\in\mathcal{D}}$, but using this information to calculate $\boldsymbol{\theta}_0$ may be too computationally demanding to be feasible (see, e.g. Ma et al., 2015; Drovandi et al., 2017; Wang et al., 2018; Deldossi and Tommasi, 2022). Any case in between those extremes may be encountered in practice. The auxiliary variables may be weakly, strongly, or even perfectly correlated with the unobserved study variables. The stronger the correlation, the greater the potential benefits of optimal sampling.

The algorithms presented in Section 3.3 for finding optimal sampling schemes require information about the full-data Hessian matrix $\mathbf{H}(\boldsymbol{\theta})$ and gradients $\boldsymbol{\psi}_i(\boldsymbol{\theta})$, evaluated at the full-data parameter $\boldsymbol{\theta}_0$. Moreover, the Hessian depends on the explanatory variables $\boldsymbol{x}_i$, if such are included in the model, and sometimes also on the outcomes $\boldsymbol{y}_i$. Similarly, the gradients depend on both the outcomes and the explanatory variables. To handle this 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 $\tilde{\boldsymbol{\theta}}_0$ of the full-data parameter $\boldsymbol{\theta}_0$, 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 Section 5.2 we will discuss how such information can be acquired gradually during the subsampling process. Below we define the anticipated covariance matrix as the target of optimisation under an assisting auxiliary model for the unknowns.

Definition 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 $\tilde{\boldsymbol{\theta}}_0$ of the full-data parameter $\boldsymbol{\theta}_0$, 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 $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ is defined as*

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

The anticipated covariance matrix in Definition 3 is our prediction of the actual unknown covariance matrix $\boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$, given the available auxiliary information. We use the term *anticipated* rather than *expected*, as adopted from Isaki and Fuller (1982), to emphasise that the expectation involved in the above definition is a hypothetical construct and generally differs from the expectation under the data generating mechanism.

All results in Section 3 and 4 may now be restated for Φ-optimality with respect to the anticipated covariance matrix $\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\mu}; \tilde{\boldsymbol{\theta}}_0)$ instead of the approximate covariance matrix $\boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$. 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 2 are replaced by their corresponding expectations

$$\tilde{c}_i := \mathrm{E}_{(\boldsymbol{x},\boldsymbol{y})\sim f(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})}\left[C_i|\{\boldsymbol{z}_i\}_{i\in\mathcal{D}}\right], \quad C_i = \left|\left|\mathbf{L}_t^{\mathsf{T}}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\right|\right|^2_{2,\boldsymbol{\theta}_0=\tilde{\boldsymbol{\theta}}_0}, \tag{17}$$

where 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 $\mathbf{L}_t\mathbf{L}_t^{\mathsf{T}} = \boldsymbol{\phi}(\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\mu}_{t-1}; \tilde{\boldsymbol{\theta}}_0))$.

We note that C_i in (17) is a positive random variable, which implies that $\tilde{c}_i > 0$ as long as $C_i > 0$ with positive probability. This is fulfilled whenever the covariance matrices for the components of $(\mathbf{X}_i, \mathbf{Y}_i)$ not included in $\boldsymbol{z}_i$ are of full rank for all i. Hence, considering the anticipated covariance under an auxiliary distribution that properly acknowledge the uncertainty in the unknowns, we effectively avoid the situation where the presented algorithms (Algorithm 1 and 2) converge to an unfeasible solution.

5.2 Sequential optimal design

The anticipated covariance introduced in the previous section takes us one step closer to a practical framework for optimal subsampling. With this notion, optimal sampling schemes may be found using the methods of Section 3.3, with the unknown values of the coefficients c_i replaced by their expectations (17) under an assisting auxiliary model $f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z})$ and a preliminary parameter estimate $\tilde{\boldsymbol{\theta}}_0$. In most cases, however, even this information is unavailable before any data is observed. This problem may be approached using sequential optimal design. Hence, subsampling is performed in multiple stages, where the information acquired from previous sampling stages may be utilised to devise optimal sampling schemes in succeeding stages. We acknowledge that many algorithms and methods in this spirit have already been presented (see, e.g. Bach, 2007; Wang et al., 2018; Imberg et al., 2020; Ai et al., 2021b). A general procedure is presented in Algorithm 3.

The number of sampling stages K in Algorithm 3 may range from a single stage with n observations, to n stages with a single observation in each subsample. In linear regression, for instance, there is no need for sequential subsampling if the explanatory variables $\boldsymbol{x}_i$ are known. This holds since in this case (17) is a function of the predictors $\boldsymbol{x}_i$ (which are known), the Hessian $\mathbf{H}(\boldsymbol{\theta}_0)$ (which only depends on the predictors $\boldsymbol{x}_i$), and the second moments of the residuals. See

Algorithm 3. K-stage subsampling procedure.

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 $\{n_k\}_{k=1}^{K}$.

```
for $k = 1, 2, \ldots, K$ do
    Calculate (optimal) sampling scheme.
    Select a random subsample of size $n_k$.
    Estimate the target parameter $\boldsymbol{\theta}_0$.
    Update the auxiliary model $f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z})$.
    Evaluate performance/precision.
    STOP if sufficient precision in reached. ELSE continue.
end for
```

Ma et al. (2020) for various optimality criteria and corresponding optimal sampling schemes in this context. At the other extreme, active learning methods utilise a large number of sampling stages, often with a single observation per stage to gain maximal flexibility in the sampling process (Bach, 2007; Imberg et al., 2020; Kossen et al., 2022; Zhan et al., 2022). Subsampling methods in big data often rely on two sampling stages: an initial simple random sample followed by an optimal unequal probability sample (Wang et al., 2018; Ai et al., 2021b; Wang and Ma, 2021).

An estimator for $\boldsymbol{\theta}_0$ after k sampling stages may be defined as

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}^{(k)} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Omega}} \hat{\ell}_{\boldsymbol{\mu}}^{(k)}(\boldsymbol{\theta}), \tag{18}$$

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

where S_{ji} is the number of times an instance $i \in \mathcal{D}$ is selected by the sampling mechanism at stage j, μ_{ji} the corresponding expected number of selections, $m_k = n_1 + \ldots + n_k$ the cumulative sample size after k stages, and $w_{ji} = 1/\mu_{ji}$. Here $\hat{\ell}_{\boldsymbol{\mu},j}(\boldsymbol{\theta})$ is an unbiased Hansen-Hurwitz estimator of the full-data empirical risk $\ell_0(\boldsymbol{\theta})$ from the sample obtained at stage j, and $\hat{\ell}_{\boldsymbol{\mu}}^{(k)}(\boldsymbol{\theta})$ a pooled estimator calculated from the first k subsamples.

The properties of the resulting estimator (18), have been studied in some specific cases, where it has been proven that under suitable regularity conditions the estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}^{(k)}$ is asymptotically normally distributed and consistent for $\boldsymbol{\theta}_0$. See, e.g., Ai et al. (2021b) and Yu et al. (2022) for results on generalised linear models and quasi-likelihood methods when the number of sampling stages $K = 2$. Imberg et al. (2022b) established the asymptotic properties of estimators for finite population vector characteristics when the subsample sizes n_k are bounded and the number of sampling stages $K \to \infty$. 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 (18) when the batch sizes n_k are bounded and $K \to \infty$ may also be deduced (cf. Zhang et al., 2021). We conjecture that a similar result holds also in the case when the number of sample stages K is bounded and the subsample sizes n_k tend to infinity, along with $N \to \infty$ and $n_k/N \to \gamma_k \in (0, 1)$. A thorough treatment of this issue, however, is a topic for future research.

6 Application and Examples

There is already an extensive amount of publications demonstrating the benefits of optimal subsampling; see, e.g., the references in Section 1. We will not provide further evidence for these already convincing results. Instead, in this section we illustrate the presented methodology through examples, and compare different optimality criteria for data subsampling in terms of computation aspects and estimator efficiency.

We consider an application in scenario generation for virtual safety assessment of an advanced driver assistance system. A brief background to the application, description of the data and problem formulation is provided in Section 6.1. Examples, illustrations and results for parametric density estimation are presented in Section 6.2, regression modelling in Section 6.3, and finite population inference in Section 6.4.

6.1 Materials and methods

Background Road traffic injuries is a major cause of death worldwide (World Health Organization, 2018). Countermeasures, such as advanced driver assistance systems, are constantly developed to mitigate these risks. One way to evaluate such systems before they enter the market is through virtual simulations (Anderson et al., 2013; Seyedi et al., 2021). Since such evaluations are performed in a virtual rather than physical test environment, they are more cost-efficient than traditional test beds. This, however, comes at the cost of a huge computational load. Computation demands can be substantially reduced through subsampling (Mullins et al., 2018; Imberg et al., 2022b; Sun et al., 2022).

Dataset Our dataset consists of 44,220 observations generated through variations of 44 reconstructed real rear-end crashes. The variations were generated by altering the driver behaviour of the ensuing vehicle in terms of glance behaviour (off-road glance duration after a specific anchoring point in time) and braking profile (maximal deceleration during braking). For each such variation, a corresponding scenario was setup in a virtual environment and simulation software, through which the entire course of events could be simulated. The outcomes of such a simulation include whether a collision occurred or not, and the impact speed if there was a collision. Thus, each observation in the dataset represents a synthetic event that describes what could have happened in the original crash event under certain variations of the conditions. Each scenario was further run under two 'treatment conditions': a scenario with an advanced emergency braking (AEB) system, and a baseline manual driving scenario without the AEB.

The following variables are included in the dataset:

- Input variables: case identifier (categorical with 44 levels corresponding to the 44 original rear-end crashes) off-road glance duration (67 levels, 0–6.6 s), and maximal deceleration during braking (15 levels, 3.3–10.3 m/s^2).
- Direct outcomes: crash indicator (1 if there was a collision and 0 otherwise) and impact speed with the AEB system and under the baseline manual driving scenario.
- Calculated outcomes: injury risk with the AEB system and under the baseline manual driving scenario, impact speed reduction, injury risk reduction, and crash avoidance indicator with the AEB system compared to baseline manual driving.

Associated with each observation is also an observation weight $w_i > 0$, describing the probability of the specific input parameter configuration (i.e., off-road glance duration and maximal deceleration during braking) occurring in real life. Additional details may be found in Imberg et al. (2022b).

Target characteristics We are interested in the following:

i) The impact speed distribution under the baseline scenario, restricted to the subset of input values that produce a crash.

ii) The impact speed response surface under the baseline scenario, as a function of the off-road glance duration and maximal deceleration.

iii) The mean impact speed reduction, mean injury risk reduction, and crash avoidance rate with the AEB compared to baseline manual driving, restricted to the subset of variations for which there is a crash in the baseline scenario.

Characteristics of the dataset, including the baseline impact speed distribution, impact speed response surface, and safety benefit distribution of the AEB compared to baseline manual driving, are presented graphically in Figure S1 and S2 in Appendix B.

As often is the case in practice, we assume that running all simulations of interest is practically unfeasible and subsampling inevitable. In such a case, the input variables (i.e., case identifier, off-road glance duration, and maximal deceleration during braking) and scenario probabilities are available *a priori* for all instances in the dataset. Hence, these are our auxiliary variables. The remaining variables can only be observed for a subset on which inference will be based. For simplicity, we restrict our consideration in problem i) (Section 6.2) and iii) (Section 6.4) to simulations that produce a crash in the baseline scenario. Thus, the 4299 observations that did not result in a crash are excluded from the corresponding evaluations.

Performance evaluation We evaluate the performance of the proposed optimal subsampling methods in terms of computation time and statistical efficiency on the application and inference problems described above. Also, for non-linear optimality criteria, we evaluate the number of iterations needed for convergence of the fixed-point iteration algorithm (Algorithm 2, Section 3.3), i.e., the time it takes to find the optimal sampling scheme. For a sampling scheme $\boldsymbol{\mu}$, the statistical efficiency of the estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ with respect to a criterion Φ is measured by the relative Φ-efficiency

$$\Phi\text{-eff}(\boldsymbol{\mu}) = \Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu}^*; \boldsymbol{\theta}_0))/\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)),$$

where $\boldsymbol{\mu}^*$ is the Φ-optimal sampling scheme (Atkinson and Donev, 1992; Pukelsheim, 1993). The relative Φ-efficiency measures the extent to which the sampling scheme $\boldsymbol{\mu}$ exhausts the maximum information for $\boldsymbol{\theta}_0$ with respect to the criterion Φ. Its inverse is the relative increase in the sample size needed to reach the same level of performance as the optimal design with respect to the Φ-optimality criterion. The relative efficiencies are evaluated analytically using the expression (7) for the approximate covariance matrix.

The following optimality criteria are considered: A-, c-, D-, and E-optimality, Φ_q-optimality with $q = 0.5, q = 5$, and $q = 10$, d_{ER}-optimality (i.e., L-optimality with $\mathbf{L} = \mathbf{H}(\boldsymbol{\theta}_0)^{1/2}$, which for all models in this evaluation also is equivalent to d_{KL}-optimality), and d_{S}-optimality (i.e., L-optimality with $\mathbf{L} = \mathbf{H}(\boldsymbol{\theta}_0)\mathbf{V}(\boldsymbol{\theta}_0)^{-1/2}$). See Section 2.3 and 4.2 for additional details and definitions. For the D-optimality criterion, the non-logarithmic version of the objective function $(\det(\boldsymbol{\Gamma})^{1/p})$ is used (Table 1, Section 2.3).

All algorithms and evaluations are implemented using the `R` language and environment for statistical computing, version 4.2.3 (R Core Team, 2023). Computations are carried out using a single core on a desktop running Windows 11 with an 2.1 GHz Intel i7 processor. The subsample size is set to 1% of the full-data size. Sampling schemes for linear optimality criteria are calculated according Algorithm 1, and sampling schemes for non-linear optimality criteria are calculated according to Algorithm 2 with tolerance parameter $\epsilon = 0.001$. The full data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i \in \mathcal{D}}$ and full-data parameter $\boldsymbol{\theta}_0$ are assumed to be known, so that the theoretically optimal sampling schemes can be found. The dataset and `R` code is available online at `https://github.com/imbhe/OSD`.

Results are presented for PO-WR and multinomial sampling designs, which produce identical analytical results. By similar means, analogous results may be obtained for PO-WOR.

6.2 Parametric density estimation

First we consider the distribution of the impact speed under the baseline scenario, illustrated in Figure S1 in Appendix B.

Model The impact speed is assumed to follow a log-normal distribution with parameter (η, σ) for the mean and standard deviation of the log impact speed. The full-data parameter $\boldsymbol{\theta}_0 = (\eta_0, \sigma_0)^\intercal$ is defined as

$$\boldsymbol{\theta}_0 = \underset{\eta \in \mathbb{R}, \sigma \in \mathbb{R}_{>0}}{\arg\min} \; \frac{1}{2} \sum_{i=1}^{N} w_i \left(\frac{(\log y_i - \eta)^2}{\sigma^2} + \log \sigma^2 \right), \tag{19}$$

where w_i is an observation weight known *a priori*, and y_i is the impact speed in scenario $i \in \mathcal{D}$. Without loss of generality, we assume that the observation weights have been normalised so that $\sum_{i=1}^{N} w_i = 1$.

Optimal sampling schemes As an illustrative example we consider the c-optimality criterion with $\mathbf{c} = (1, 0)^\intercal$, i.e., minimising the variance of estimating the location parameter η_0. Since the optimality criterion is linear here, the optimal sampling scheme can be found according to Algorithm 1 with

$$\mathbf{L} = (1, 0)^\intercal, \quad \mathbf{H}(\boldsymbol{\theta}_0) = \frac{1}{\sigma_0^2} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \quad \boldsymbol{\psi}_i(\boldsymbol{\theta}_0) = -w_i \left(\frac{\log y_i - \eta_0}{\sigma_0^2}, \frac{(\log y_i - \eta_0)^2}{\sigma_0^3} - \frac{1}{\sigma_0} \right)^\intercal,$$

and

$$c_i = \left|\left| \mathbf{L}^\intercal \mathbf{H}(\boldsymbol{\theta}_0)^{-1} \boldsymbol{\psi}_i(\boldsymbol{\theta}_0) \right|\right|_2^2 \propto w_i^2 (\log y_i - \eta_0)^2. \tag{20}$$

To find an optimal sampling scheme with respect to the anticipated variance of our estimator for η_0, we replace y_i by a random variable Y_i and evaluate the corresponding expectation of (20) under an assumed model for Y_i. If we assume that $\log Y_i$ has mean $\hat{y}_i$ and variance σ_i^2, we obtain

$$\tilde{c}_i \propto \sqrt{\mathrm{E}_{Y_i}[w_i^2 (\log Y_i - \eta_0)^2]} = w_i \sqrt{(\hat{y}_i - \eta_0)^2 + \sigma_i^2},$$

which in practice may be evaluated at a preliminary estimate $\tilde{\eta}_0$ of η_0. The predictions $\hat{y}_i$ and dispersion parameters σ_i^2 may be modelled as functions of the observed auxiliary variables (i.e., the case identifier, off-road glance duration, and maximal deceleration during braking), and estimated from a pilot sample or using sequential subsampling methods (Algorithm 3). The resulting sampling scheme is guaranteed to produce strictly positive sampling probabilities as long as all $w_i, \sigma_i > 0$.

Results The computation time, number of iterations needed for convergence, and relative efficiencies for various optimality criteria are presented in Table 2. The D-optimal sampling scheme was found in four fixed-point iterations with Algorithm 2. The $\Phi_{0.5}$-, Φ_5- optimal sampling schemes were found in three and 25 iterations, respectively. An E-optimal sampling scheme could not be found, due to the non-convexity of the objective function. The computation time for finding an optimal sampling scheme ranged from 0.10 seconds for the linear optimality criteria to 0.96 s for the D-optimality criterion and 4.95 s for the Φ_5-optimality criterion. The optimal sampling schemes of the d_{ER}- and d_{S}-optimality criteria reached 97–99% A-efficiency, 96–99% D-efficiency, and 92–94% Φ_5-efficiency. The A-optimal sampling scheme had a similar performance.

Table 2. Performance measures for estimating the log-normal model (19) by optimal subsampling with various optimality criteria. The columns show the number of fixed-point iterations and execution time to find the optimal sampling scheme, and relative efficiencies with respect to other optimality criteria.

Optimality criterion	No. iterations	Time (s)	A-eff	$\mathrm{c}_{(1,0)}$-eff	$\mathrm{c}_{(0,1)}$-eff	D-eff	d_{ER}-eff	Φ_5-eff
A		0.11	**1.00**	0.90	0.76	0.99	0.99	0.96
c, $\mathbf{c} = (1,0)^{\mathsf{T}}$		0.10	0.14	**1.00**	0.04	0.24	0.10	0.09
c, $\mathbf{c} = (0,1)^{\mathsf{T}}$		0.10	0.27	0.16	**1.00**	0.47	0.34	0.19
D	4	0.96	0.98	0.86	0.79	**1.00**	0.99	0.91
d_{ER}		0.10	0.99	0.84	0.83	0.99	**1.00**	0.92
d_{S}		0.10	0.97	0.84	0.79	0.96	0.98	0.94
E	Diverged	-						
$\Phi_{0.5}$	3	0.78	>0.99	0.88	0.78	>0.99	0.99	0.94
Φ_5	25	4.95	0.95	0.91	0.65	0.90	0.91	**1.00**
Φ_{10}	Diverged	-						

6.3 Regression modelling

Next we consider the distribution of the baseline impact speed as a function of the input variables to the scenario generation, i.e., the off-road glance duration and maximal deceleration during braking.

We first note that the impact speed increases monotonically with increased levels of the off-road glance duration and decreased levels of deceleration. Hence, variations generated from the same original rear-end crash have an upper bound on their impact speed, attained for the variation having the off-road glance duration at its maximum and the deceleration level at its minimum. We assume that this maximal impact speed is known, e.g., observed by running the corresponding virtual simulation. The impact speed may then be expressed relative to the maximal impact speed for that specific case, with values in the common range $[0, 1]$. Note that in this case the explanatory variables are known *a priori*, whereas the outcome (i.e., relative impact speed) can only be observed after running the corresponding virtual simulation.

Model A simple model for a response variable on the unit interval is a quasi-binomial logistic regression model, for which the full-data parameter $\boldsymbol{\theta}_0$ is defined as

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\arg\min} - \sum_{i=1}^{N} y_i \log p_i(\boldsymbol{\theta}) + (1-y_i)\log(1-p_i(\boldsymbol{\theta})), \quad p_i(\boldsymbol{\theta}) = (1+\exp(-\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\theta}))^{-1}, \tag{21}$$

where $\boldsymbol{x}_i$ is a feature vector pertaining to instance i, and $\boldsymbol{\theta}$ a vector of regression coefficients. As explanatory variables we include the case identifier of the original rear-end crash event (categorical with 44 levels, dummy coded into 44 binary variables), the off-road glance duration, the maximal deceleration during braking, and all three-way interactions. For each of the 44 cases, the impact speed response surface is then described by 4 parameters: an intercept parameter and three slope parameters corresponding to the off-road glance duration, deceleration level, and the interaction between those. The joint parameter vector $\boldsymbol{\theta}$ is of dimension $44 \times 4 = 176$. Note that in this case we do not include the observation weights w_i in the empirical risk function, since these are functions of the explanatory variables and hence ignorable in this context. Illustrations of the observed and predicted impact speed response surfaces for three of the cases are presented in Figure S2 in Appendix B.

Optimal sampling schemes For illustrative purposes, we consider the d_{ER}-optimality criterion. Since this is a linear optimality criterion, the optimal sampling scheme can be found according to Algorithm 1 with

$$\mathbf{L} = \mathbf{H}(\boldsymbol{\theta}_0)^{1/2}, \quad \mathbf{H}(\boldsymbol{\theta}_0) = \mathbf{X}^{\mathsf{T}}\mathbf{W}(\boldsymbol{\theta}_0)\mathbf{X}, \quad \boldsymbol{\psi}_i(\boldsymbol{\theta}_0) = -(y_i - p_i(\boldsymbol{\theta}_0))\boldsymbol{x}_i,$$

where $\mathbf{W}(\boldsymbol{\theta})$ is the diagonal matrix with entries $p_i(\boldsymbol{\theta})(1-p_i(\boldsymbol{\theta}))$ and $\mathbf{X}$ the matrix with rows $\boldsymbol{x}_i^\mathsf{T}$, and

$$c_i = ||\mathbf{L}^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)||_2^2 = (y_i - p_i(\boldsymbol{\theta}_0))^2\boldsymbol{x}_i^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{W}(\boldsymbol{\theta}_0)\mathbf{X})^{-1}\boldsymbol{x}_i. \tag{22}$$

To find a d_{ER}-optimal sampling scheme with respect to the anticipated covariance matrix, we replace y_i in (22) by a random variable Y_i and evaluate the corresponding expectation under a model for Y_i given the known explanatory variables $\boldsymbol{x}_i$. For instance, we may assume that Y_i has mean $p(\boldsymbol{\theta}_0)$ and variance $p_i(\boldsymbol{\theta}_0)(1-p_i(\boldsymbol{\theta}_0))$. We then obtain

$$\tilde{c}_i = \sqrt{\mathrm{E}_{Y_i}[(Y_i - p_i(\boldsymbol{\theta}_0))^2\boldsymbol{x}_i^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{W}(\boldsymbol{\theta}_0)\mathbf{X})^{-1}\boldsymbol{x}_i]} = \sqrt{h_{ii}(\boldsymbol{\theta}_0)}, \tag{23}$$

where $h_{ii}(\boldsymbol{\theta}_0)$ is the i^{th} diagonal element of the 'hat matrix', or projection matrix

$$\mathbf{W}(\boldsymbol{\theta}_0)^{1/2}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{W}(\boldsymbol{\theta}_0)\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}(\boldsymbol{\theta}_0)^{1/2}$$

(see Hoaglin and Welsch, 1978; Pregibon, 1981). To account for the influence a data point $(\boldsymbol{x}_i, y_i)$ exerts on its own prediction, it is appropriate to deflate the variance of Y_i by a factor $1 - h_{ii}(\boldsymbol{\theta}_0)$, resulting in

$$\tilde{c}_i = \sqrt{h_{ii}(\boldsymbol{\theta}_0)(1-h_{ii}(\boldsymbol{\theta}_0))}$$

instead of (23) (cf. Ma et al., 2020). In practice we may evaluate $\tilde{c}_i$ at a preliminary estimate $\tilde{\boldsymbol{\theta}}_0$ obtained from a pilot sample or estimated using sequential subsampling methods (Algorithm 3). The resulting sampling scheme is guaranteed to produce strictly positive sampling probabilities as long as the predictions $p_i(\tilde{\boldsymbol{\theta}}_0)$ are bounded away from 0 and 1.

Results Table 3 shows the computation time, relative efficiencies, and number of iterations needed to find an optimal sampling scheme for various optimality criteria. Optimal sampling schemes were found in five fixed-point iterations for D-optimality, four iterations for $\Phi_{0.5}$-optimality, and could not be found for the Φ_5-, Φ_{10}- and E-optimality criteria. Finding an L-optimal sampling scheme required 95% less computation time than for the non-linear D-optimality criterion. The d_{ER}- and d_{S}-optimal schemes attained 40–47% A-efficiency and 92–96% D-efficiency. The A-optimal sampling scheme had only 60% D-efficiency. The $\Phi_{0.5}$-optimal sampling scheme, which interpolates between A- and D-optimality, achieved 92% A-efficiency and 80% D-efficiency.

Table 3. Performance measures for estimating the quasi-binomial logistic regression model (21) by optimal subsampling with various optimality criteria. The columns show the number of fixed-point iterations and execution time to find the optimal sampling scheme, and relative efficiencies with respect to other optimality criteria. The computation time for fitting the model to the full dataset was 8.46 seconds.

Optimality criterion	No. iterations	Time (s)	A-eff	D-eff	d_{ER}-eff	d_{S}-eff	$\Phi_{0.5}$-eff
A		1.16	**1.00**	0.60	0.47	0.42	0.93
D	5	27.69	0.49	**1.00**	0.89	0.94	0.77
d_{ER}		1.11	0.47	0.92	**1.00**	0.91	0.71
d_{S}		1.12	0.40	0.96	0.92	**1.00**	0.67
E	Diverged	-					
$\Phi_{0.5}$	4	20.37	0.92	0.80	0.68	0.65	**1.00**
Φ_5	Diverged	-					
Φ_{10}	Diverged	-					

6.4 Finite population inference

We finally consider the potential safety benefit of the AEB system compared to a baseline manual driving scenario. For a scenario $i \in \mathcal{D}$, let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, y_{i3})^\mathsf{T}$, where y_{1i} is the impact speed reduction, y_{i2} the injury risk reduction, and y_{i3} the binary crash avoidance indicator with the AEB system compared to baseline manual driving. The distributions of these characteristics are illustrated in Figure S1 in Appendix B.

Model We are interested in the mean impact speed reduction, mean injury risk reduction and crash avoidance rate, given by the vector total

$$\boldsymbol{t}_{\boldsymbol{y}} = \sum_{i=1}^{N} w_i\boldsymbol{y}_i,$$

where the observation weights w_i are normalised so that $\sum_{i=1}^{N} w_i = 1$. This can also be expressed as

$$\boldsymbol{t}_{\boldsymbol{y}} = \boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\mathbb{R}^3}{\arg\min}\,\frac{1}{2}\sum_{i=1}^{N} w_i||\boldsymbol{y}_i - \boldsymbol{\theta}||_2^2. \tag{24}$$

Optimal sampling schemes As an example, consider the d_{S}-optimality criterion. Since this is a linear optimality criterion, the optimal sampling scheme can be found according to Algorithm 1 with

$$\mathbf{L} = \mathbf{H}(\boldsymbol{\theta}_0)\mathbf{V}(\boldsymbol{\theta}_0)^{-1/2}, \quad \mathbf{H}(\boldsymbol{\theta}_0) = \mathbf{I}_{3\times 3}, \quad \mathbf{V}(\boldsymbol{\theta}_0) = \sum_{i\in\mathcal{D}} \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^T, \quad \boldsymbol{\psi}_i(\boldsymbol{\theta}_0) = -w_i(\boldsymbol{y}_i - \boldsymbol{\theta}_0),$$

and

$$c_i = ||\mathbf{L}^{\mathsf{T}}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)||_2^2 = w_i^2(\boldsymbol{y}_i - \boldsymbol{\theta}_0)^{\mathsf{T}}\mathbf{V}(\boldsymbol{\theta}_0)^{-1}(\boldsymbol{y}_i - \boldsymbol{\theta}_0). \tag{25}$$

In order to find the L-optimal sampling scheme with respect to the anticipated covariance matrix, we introduce a random vector $\mathbf{Y}_i$, substitute $\mathbf{Y}_i$ for $\boldsymbol{y}_i$ in (25), and evaluate the expectation. Let therefore $\hat{\boldsymbol{y}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$ denote the mean vector and covariance matrix of $\mathbf{Y}_i$, respectively. By properties of quadratic forms (Mathai and Provost, 1992), we obtain

$$\tilde{c}_i = \sqrt{\mathrm{E}\left[w_i^2(\boldsymbol{y}_i - \boldsymbol{\theta}_0)^{\mathsf{T}}\mathbf{V}(\boldsymbol{\theta}_0)^{-1}(\boldsymbol{y}_i - \boldsymbol{\theta}_0)\right]} = w_i\sqrt{\left[(\hat{\boldsymbol{y}}_i - \boldsymbol{\theta}_0)^{\mathsf{T}}\mathbf{V}(\boldsymbol{\theta}_0)^{-1}(\hat{\boldsymbol{y}}_i - \boldsymbol{\theta}_0) + \mathrm{tr}(\mathbf{V}(\boldsymbol{\theta}_0)^{-1}\hat{\boldsymbol{\Sigma}}_i)\right]}. \tag{26}$$

To implement optimal sampling in practice, we evaluate (26) at a preliminary estimate $\tilde{\boldsymbol{\theta}}_0$ obtained from a pilot sample. The predictions $\hat{\boldsymbol{y}}_i$ and dispersion matrices $\hat{\boldsymbol{\Sigma}}_i$ may be modelled as functions of the observed auxiliary variables (i.e., the case identifier, off-road glance duration, and maximal deceleration during braking) and iteratively updated using sequential subsampling methods (Algorithm 3). The sampling scheme derived from (26) is guaranteed to produce strictly positive sampling probabilities as long as all $\hat{\boldsymbol{\Sigma}}_i$ are full-rank.

Results Results in terms of computation time, number of iterations needed for convergence, and relative efficiencies of various optimality criteria are presented in Table 4. The optimal sampling scheme was found in four fixed-point iterations for the D-optimality criterion, and in two iterations for the other non-linear optimality criteria. The computation time ranged from 0.10 for the linear optimality criteria, to 0.94 s for the D-optimality criterion. The d_{ER}-optimal sampling scheme had 100% A-efficiency, 46% D-efficiency and >99% E-efficiency. In fact, in this case the d_{ER}-optimality criterion is identical to A-optimality. In contrast, the d_{S}–optimal sampling scheme had 73% A-efficiency, 98% D-efficiency, and 72% E-efficiency. The $\Phi_{0.5}$-criterion had 99% A-efficiency, 58% D-efficiency and 99% E-efficiency. The A- and E-optimality criteria were largely driven by the mean impact speed reduction, as this was measured on a scale that was orders of magnitude larger than the measurement-scale for the injury risk reduction and crash avoidance (Figure S1, Appendix B).

Table 4. Performance measures for estimating the vector of finite population means (24) by optimal subsampling with various optimality criteria. The columns show the number of fixed-point iterations and execution time to find the optimal sampling scheme, and relative efficiencies with respect to other optimality criteria.

Optimality criterion	No. iterations	Time (s)	A-eff	$\mathrm{c}_{(1,0,0)}$-eff	$\mathrm{c}_{(0,1,0)}$-eff	$\mathrm{c}_{(0,0,1)}$-eff	D-eff	E-eff
A		0.10	**1.00**	>0.99	0.36	0.25	0.46	>0.99
c, $\mathbf{c} = (1,0,0)^{\mathsf{T}}$		0.10	0.98	**1.00**	0.05	0.04	0.20	>0.99
c, $\mathbf{c} = (0,1,0)^{\mathsf{T}}$		0.10	0.12	0.12	**1.00**	0.11	0.22	0.12
c, $\mathbf{c} = (0,0,1)^{\mathsf{T}}$		0.10	0.41	0.41	0.50	**1.00**	0.70	0.41
D	4	0.94	0.65	0.65	0.76	0.82	**1.00**	0.65
d_{ER}		0.10	**1.00**	>0.99	0.36	0.25	0.46	>0.99
d_{S}		0.10	0.73	0.72	0.77	0.77	0.98	0.72
E	2	0.70	0.99	>0.99	0.07	0.06	0.22	**1.00**
$\Phi_{0.5}$	2	0.57	0.99	0.99	0.46	0.38	0.58	0.99
Φ_5	2	0.58	0.99	>0.99	0.07	0.06	0.22	**1.00**
Φ_{10}	2	0.57	0.99	>0.99	0.07	0.06	0.22	**1.00**

7 Discussion

We have presented a theory of optimal subsampling design for a general class of estimators, sampling designs, and optimality criteria. Although the presented optimality conditions are valid for any differentiable objective function, the algorithms for finding optimal sampling schemes are most appropriate for convex functions. Further research could include development of methods to handle non-convex optimality criteria, such as E- and G-optimality (Kiefer and Wolfowitz, 1960; Kiefer, 1974).

From an applied perspective, we believe that the proposed invariant linear optimality criteria (i.e., d_{ER}-, d_{KL}- and d_{S}-optimality) offer a good compromise between computational and statistical efficiency. Non-linear optimality criteria require iterative procedures and computationally expensive covariance matrix evaluations, which limits their usability in problems and applications where computational complexity is a major concern. Further studies evaluating the performance of these methods in practice and in other applications are encouraged.

Sequential subsampling is a viable approach to implement optimal subsampling methods in practice. The theoretical properties of the estimators derived from such sequential subsampling methods have so far only been studied rigorously in limited settings. Further research in this direction is requested.

Acknowledgement

We would like to thank Malin Svärd and Simon Lundell at Volvo Car Corporation for allowing us to use their data in our experiments.

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*.

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*.

Bellhouse, D. R. (1984). A review of optimal designs in survey sampling. *Canadian Journal of Statistics*, 12(1):53–65.

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

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

Brewer, K. R. W. (1979). A class of robust sampling designs for large-scale surveys. *Journal of the American Statistical Association*, 74(368):911–915.

Casella, G. and Berger, R. (2001). *Statistical Inference*. Duxbury, Pacific Grove.

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.

Dai, W., Song, Y., and Wang, D. (2022). A subsampling method for regression problems based on minimum energy criterion. *Technometrics*. Advance online publication. `https://doi.org/10.1080/00401706.2022.2127915`.

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

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.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65(3):457–482.

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

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

Hall, P. and Heyde, 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.

Hastie, T. (2020). Ridge regularization: An essential concept in data science. *Technometrics*, 62(4):426–433.

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

Horn, R. and Johnson, C. (1990). *Matrix Analysis*. Cambridge University Press, Cambridge.

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.

Hájek, J. (1959). Optimal strategy and other problems in probability sampling. *Časopis pro pěstování matematiky*, 84(4):387–423.

Imberg, H., Jonasson, J., and Axelson-Fisk, M. (2020). Optimal sampling in unbiased active learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*.

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

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

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.

Kiefer, J. (1974). General Equivalence Theory for Optimum Designs (Approximate Theory). *The Annals of Statistics*, 2(5):849–879.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366.

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*.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

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

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*.

Mathai, A. and Provost, S. (1992). *Quadratic Forms in Random Variables*. Marcel Dekker, New York.

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

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 (General)*, 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*.

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

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.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

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. `https://doi.org/10.1155/2021/4430730`.

Sibson, R. (1974). D_A-optimality and duality. In *Progress of Statistics, Volume 2: Proceedings of the 9th European Meeting of Statisticians*.

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

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

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.

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

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

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

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, Y., Yu, A. W., and Singh, A. (2017). On computationally tractable selection of experiments in measurement-constrained regression models. *Journal of Machine Learning Research*, 18(143):1–41.

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.

World Health Organization (2018). *Global status report on road safety 2018*. `https://www.who.int/publications/i/item/9789241565684`.

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

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.

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.

A Proofs

A.1 Proof of Lemma 1

According to the chain rule in matrix differential calculus (Petersen and Pedersen, 2012) we have that

$$\frac{\partial \Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial \mu_i} = \operatorname{tr}\left(\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))^{\mathsf{T}} \frac{\partial \boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)}{\partial \mu_i}\right),$$

where $\boldsymbol{\phi}(\mathbf{U}) = \frac{\partial \Phi(\mathbf{U})}{\partial \mathbf{U}}$ is the $p \times p$ matrix derivative of Φ with respect to its matrix argument, and $\frac{\partial \boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)}{\partial \mu_i}$ the elementwise derivative of $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ with respect to μ_i. Since $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ is symmetric, $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ must also be symmetric, which proves (11).

By the assumptions, $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ decreases monotonically with μ_i in the Loewner order sense, which implies that $\frac{\partial \boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)}{\partial \mu_i}$ is negative semi-definite. Therefore, there exists a real matrix $\mathbf{U}$ such that $\mathbf{U}\mathbf{U}^{\mathsf{T}} = -\frac{\partial \boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)}{\partial \mu_i}$. Moreover, $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ is monotone for Loewner's ordering and hence a monotone decreasing function of μ_i, so we must have

$$\frac{\partial \Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial \mu_i} \le 0,$$

which by the above is equivalent to

$$\operatorname{tr}\left(\mathbf{U}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))\mathbf{U}\right) \ge 0, \quad \mathbf{U}\mathbf{U}^{\mathsf{T}} = -\frac{\partial \boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)}{\partial \mu_i}.$$

This inequality holds true for every $\mathbf{U}$, and hence for every possible value of the matrix $\frac{\partial \boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)}{\partial \mu_i}$, if and only if $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ is positive semi-definite. Consequently, there exists a real $p \times p$ matrix $\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ such that $\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{\mathsf{T}} = \boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$.

A.2 Proof of Lemma 2

Proof of a) We have by (7) and (9) that

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

Taking the derivative with respect to μ_i, we obtain

$$\frac{\partial\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)}{\partial\mu_i} = -\mu_i^{-2}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^{\mathsf{T}}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}.$$

Proof of b)–e) Follows by the following rules from matrix differential calculus (Petersen and Pedersen, 2012):

b) $\frac{\partial \log\det(\mathbf{U})}{\partial U} = \mathbf{U}^{-1}$, provided that $\mathbf{U}$ is of full rank.

c) $\frac{\partial\lambda_{\max}(\mathbf{U})}{\partial U} = \mathbf{v}\mathbf{v}^{\mathsf{T}}$, where $\mathbf{v}$ is an eigenvector pertaining to the maximal eigenvalue of $\mathbf{U}$, provided that $\mathbf{v}$ is unique.

d) $\frac{\partial\mathrm{tr}(\mathbf{U}\boldsymbol{A})}{\partial U} = \boldsymbol{A}^{\mathsf{T}}$.

e) $\frac{\partial\mathrm{tr}(\mathbf{U}^q)}{\partial\mathbf{U}} = q(\mathbf{U}^{q-1})^{\mathsf{T}}$, so that $\frac{\partial\mathrm{tr}(\mathbf{U}^q)^{1/q}}{\partial U} = \frac{1}{q}\mathrm{tr}(\mathbf{U}^q)^{1/q-1}\frac{\partial\mathrm{tr}(\mathbf{U}^p)}{\partial\mathbf{U}} = \mathrm{tr}(\mathbf{U}^q)^{1/q-1}(\mathbf{U}^{q-1})^{\mathsf{T}}$, provided that $\mathbf{U}$ is of full rank. The final result follows by symmetry of $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$.

A.3 Proof of Lemma 3

Combining the results of Lemma 1 and 2, we observe for PO-WR, PO-WOR and MULTI designs that the partial derivative of $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ with respect to μ_i, whenever it exists, is given by

$$\begin{aligned}\frac{\partial\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i} &= -\mathrm{tr}\left(\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))\mu_i^{-2}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^{\mathsf{T}}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\right) \\ &= -\mu_i^{-2}\mathrm{tr}(\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^{\mathsf{T}}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{\mathsf{T}}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)) \\ &= -\mu_i^{-2}\left|\left|\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{\mathsf{T}}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\right|\right|_2^2.\end{aligned}$$

The second equality follows from the cyclic property of the trace and definition of $\mathbf{L}(\boldsymbol{\mu},\boldsymbol{\theta}_0)$, and the third by noting that the expression within the parentheses is a scalar and equals the squared Euclidean norm of the vector $\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{\mathsf{T}}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)$.

A.4 Proof of Lemma 4

By a second order Taylor expansion around $\boldsymbol{\theta}_0$, we have that

$$d(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}) = d(\boldsymbol{\theta}_0) + \nabla d(\boldsymbol{\theta})^{\mathsf{T}}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0)^{\mathsf{T}}\mathbf{H}_d(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0) + o_p(||(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0)||_2^2),$$

where the first two terms, by definition of $d(\boldsymbol{\theta})$, are zero, and $\mathbf{H}_d(\boldsymbol{\theta}_0) = \frac{\partial^2 d(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}}$. By the assumptions on $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$, we have that $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}} - \boldsymbol{\theta}_0 = o_p(n^{-1/2})$ and $\mathrm{E}[|\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0|^{2+\delta}] < \infty$ (elementwise) for some $\delta > 0$. By bounded convergence, this implies for the remainder that $\mathrm{E}[o_p(||(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0)||_2^2)] = o(n^{-1})$. We have further that

$$\begin{aligned}\mathrm{E}\left[(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0)^{\mathsf{T}}\mathbf{H}_d(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0)\right] &= \mathrm{tr}\left(\mathbf{H}_d(\boldsymbol{\theta}_0)\mathbf{Cov}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0)\right) + \mathrm{E}[\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0]^{\mathsf{T}}\mathbf{H}_d(\boldsymbol{\theta}_0)\mathrm{E}[\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0] \\ &= \mathrm{tr}\left(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{H}_d(\boldsymbol{\theta}_0)\right) + o(n^{-1}),\end{aligned}$$

where the first equality follows from properties of quadratic forms (Mathai and Provost, 1992), and the second by assumptions (6)–(7) on $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ and the cyclic property of the trace.

A.5 Proof of Proposition 1

First we note that the matrix $\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)$ exists by Lemma 1 whenever the objective function is differentiable at $\boldsymbol{\mu}^*$. Hence, the coefficients c_i are positive, the square roots $\sqrt{c_i}$ are real, and the optimality conditions (13) and (14a)–(14c) well-defined.

Proof of a) Consider the function $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ subject to the constraints $\sum_{i\in\mathcal{D}}\mu_i = n$, and $\mu_i > 0$ for all $i \in \mathcal{D}$. By the Lagrange multiplier method (Boyd and Vandenberghe, 2004), the constrained stationary points of $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ are obtained as the stationary points of the Lagrangian

$$\Lambda(\boldsymbol{\mu},\alpha) = \Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) + \alpha g(\boldsymbol{\mu}), \quad g(\boldsymbol{\mu}) = \sum_{i\in\mathcal{D}}\mu_i - n.$$

Taking the derivatives with respect to $\boldsymbol{\mu}$ and α, we obtain the system of equations

$$\nabla\Lambda(\boldsymbol{\mu},\alpha) = \mathbf{0} \quad \Leftrightarrow \quad \begin{cases} g(\boldsymbol{\mu}) = 0 \\ -\nabla_{\boldsymbol{\mu}}\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \alpha\nabla g(\boldsymbol{\mu}). \end{cases}$$

Now, $\frac{\partial\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i} = -c_i/\mu_i^2$ by Lemma 3 and definition of c_i, and $\frac{\partial g(\boldsymbol{\mu})}{\partial\mu_i} = 1$. A stationary point therefore satisfies the system of equations $\alpha = c_1/\mu_1^2 = \ldots = c_N/\mu_N^2$ for all $i \in \mathcal{D}$. For $\boldsymbol{\mu}^*$ to be Φ-optimal we must have $\mu_i^* \propto \sqrt{c_i}$, $\mu_i^* > 0$, and $\sum_{i\in\mathcal{D}}\mu_i = n$, and hence

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

Proof of b) Consider the function $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ subject to the constraints $\sum_{i\in\mathcal{D}}\mu_i = n$ and $0 < \mu_i \le 1$ for all $i \in \mathcal{D}$. Also consider the Lagrangian

$$\Lambda(\boldsymbol{\mu},\alpha,\boldsymbol{\beta}) = \Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) + \alpha g(\boldsymbol{\mu}) + \sum_{i\in\mathcal{D}}\beta_i h_i(\boldsymbol{\mu}),$$

where $g(\boldsymbol{\mu}) = \sum_{i\in\mathcal{D}}\mu_i - n$ and $h_i(\boldsymbol{\mu}) = \mu_i - 1$. The constrained stationary points of $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ are characterised as the solutions to the Karush-Kuhn-Tucker conditions (Boyd and Vandenberghe, 2004):

- *Stationarity*: $-\nabla_{\boldsymbol{\mu}}\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \alpha\nabla g(\boldsymbol{\mu}) + \sum_{i\in\mathcal{D}}\beta_i\nabla h_i(\boldsymbol{\mu})$.
- *Primal feasibility*: $g(\boldsymbol{\mu}) = 0$, and $h_i(\boldsymbol{\mu}) \le 0$ for all $i \in \mathcal{D}$.
- *Dual feasibility*: $\beta_i \ge 0$ for all $i \in \mathcal{D}$.
- *Complementary slackness*: $\beta_i h_i(\boldsymbol{\mu}) = 0$ for all $i \in \mathcal{D}$.

First note that $\frac{\partial\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i} = -c_i/\mu_i^2$ by Lemma 3 and definition of c_i, $\frac{\partial g(\boldsymbol{\mu})}{\partial\mu_i} = 1$, and $\frac{\partial h_i(\boldsymbol{\mu})}{\partial\mu_j} = 1$ if $i = j$ and 0 otherwise. Consider a sampling scheme $\boldsymbol{\mu} \in \mathcal{M}_n$ and let $\mathcal{E} = \{i \in \mathcal{D} : \mu_i = 1\}$ and $n_{\mathcal{E}} = |\mathcal{E}|$. For the Karush-Kuhn-Tucker conditions to be satisfied, we must have that

i) $\mu_i \le 1$ and $\sum_{i\in\mathcal{D}\setminus\mathcal{E}}\mu_i = n - n_{\mathcal{E}}$, by the primal feasibility condition,

ii) $\beta_i = 0$ if $\mu_i < 1$, by the complementary slackness condition,

iii) $c_i/\mu_i^2 = \alpha + \beta_i$ by the stationarity condition, which by the above implies that

$$c_i = \begin{cases} \alpha + \beta_i & \text{if } \mu_i = 1 \\ \alpha\mu_i^2 & \text{if } \mu_i < 1 \end{cases} \quad \Leftrightarrow \quad \begin{cases} \beta_i = c_i - \alpha & \text{if } \mu_i = 1 \\ \alpha = c_i/\mu_i^2 & \text{if } \mu_i < 1, \end{cases}$$

iv) $\beta_i \ge 0$ by the dual feasibility condition, which by the above implies that $\beta_i = c_i - \alpha = c_i - c_j/\mu_j^2 \ge 0$ for $i \in \mathcal{E}$ and $j \in \mathcal{D} \setminus \mathcal{E}$.

The condition (14a) follows from i), (14b) from i) and iii), and (14c) from iv).

A.6 Proof of Proposition 2

Note first that the domain $\mathcal{M}_n$ of $\boldsymbol{\mu}$ is convex. The results hence follow from the second derivative test by showing that the Hessian matrix of $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ is positive semi-definite on $\mathcal{M}_n$.

Proof of a) We have by Lemma 3 that

$$\frac{\partial \Phi(\mathbf{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial \mu_i} = -\mu_i^{-2}\big|\big|\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\big|\big|_2^2.$$

For the L-optimality criterion the matrix $\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0) = \mathbf{L}$ does not depend on $\boldsymbol{\mu}$. The second-order partial derivatives are given by

$$\frac{\partial^2 \Phi(\mathbf{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial \mu_i \partial \mu_j} = \begin{cases} 2\mu_i^{-3}\big|\big|\mathbf{L}^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\big|\big|_2^2 \geq 0 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

This matrix is diagonal with non-negative entries for all $\mu_i > 0$, and hence positive semi-definite on $\mathcal{M}_n$.

Proof of b) We show that $\det(\mathbf{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ is log-convex in $\boldsymbol{\mu}$, i.e., that $\log\det(\mathbf{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ is convex.

First note that $\log\det(\mathbf{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \log\det(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))) - 2\log\det(\mathbf{H}(\boldsymbol{\theta}_0))$, where $\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ is given by (9) and $\mathbf{H}(\boldsymbol{\theta}_0)$ does not depend on $\boldsymbol{\mu}$. Thus, it suffices to show that $\log\det(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ is convex in $\boldsymbol{\mu}$. We obtain the desired result by showing that the Hessian of $\Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ can be decomposed as the Hadamard product between two positive semi-definite matrices, and hence is positive semi-definite (Horn and Johnson, 1990).

Consider first a PO-WR or multinomial sampling design. The partial derivatives of $\Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ are given by

$$\begin{aligned} \frac{\partial \Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial \mu_i} &= -\mu_i^{-2}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0), \\ \frac{\partial^2 \Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial \mu_i^2} &= 2\mu_i^{-3}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0) - \mu_i^{-4}(\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0))^2, \\ \frac{\partial^2 \Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial \mu_i \partial \mu_j} &= -(\mu_i\mu_j)^{-2}(\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_j(\boldsymbol{\theta}_0))^2, \quad i \neq j. \end{aligned}$$

These results follow in analogy with the proof of Lemma 3 by the chain rule (11) and the following rules for matrix differentiation (Petersen and Pedersen, 2012):

$$\frac{\partial \mathbf{a}^\mathsf{T}\mathbf{X}\mathbf{a}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{a}^\mathsf{T}, \quad \frac{\partial \log\det(\mathbf{Y})}{\partial x} = \operatorname{tr}\left(\mathbf{Y}^{-1}\frac{\partial \mathbf{Y}}{\partial x}\right), \quad \text{and} \quad \frac{\partial \mathbf{Y}^{-1}}{\partial x} = -\mathbf{Y}^{-1}\frac{\partial \mathbf{Y}}{\partial x}\mathbf{Y}^{-1}.$$

Let $\boldsymbol{u}_i = \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)/\sqrt{\mu_i}$ and $\mathbf{U}$ be the matrix with rows $\boldsymbol{u}_i^\mathsf{T}$. Also, let $\mathbf{A} = \mathbf{U}(\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\mathbf{U}^\mathsf{T}$ and a_{ij} the elements of $\mathbf{A}$. We note the following:

- $\mathbf{U}^\mathsf{T}\mathbf{U} = \sum_{i\in\mathcal{D}} \mu_i^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T} = \mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$,
- $a_{ij} = \boldsymbol{u}_i^\mathsf{T}(\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\boldsymbol{u}_j = (\mu_i\mu_j)^{-1/2}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_j(\boldsymbol{\theta}_0)$,
- $\mathbf{A}$ is an idempotent matrix, i.e., $\mathbf{A}^2 = \mathbf{A}$, which implies that $a_{ii} = \sum_j a_{ij}^2$,
- $a_{ii} = \boldsymbol{u}_i^\mathsf{T}(\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\boldsymbol{u}_i = \boldsymbol{u}_i^\mathsf{T}\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{-1}\boldsymbol{u}_i > 0$, since $\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ by assumption is positive definite.

We may now write

$$\frac{\partial^2 \Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial \mu_i \partial \mu_j} = \begin{cases} \mu_i^{-2}(2a_{ii} - a_{ii}^2) & \text{if } i = j, \\ (\mu_i\mu_j)^{-1}a_{ij}^2 & \text{if } i \neq j. \end{cases}$$

We recognise the Hessian matrix $\frac{\partial^2 \Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial \boldsymbol{\mu}\partial \boldsymbol{\mu}^\mathsf{T}}$ as the Hadamard product $\mathbf{M} \otimes \mathbf{B}$ of a rank-one matrix $\mathbf{M} = \mathbf{m}\mathbf{m}^\mathsf{T}$ with $\mathbf{m} = (\mu_1^{-1}, \ldots, \mu_N^{-1})^\mathsf{T}$, and a symmetric matrix $\mathbf{B}$ with entries

$$b_{ij} = \begin{cases} 2a_{ii} - a_{ii}^2 & \text{if } i = j, \\ a_{ij}^2 & \text{if } i \neq j. \end{cases}$$

The matrix $\mathbf{M}$ has eigenvalues $\mathbf{m}^\mathsf{T}\mathbf{m}$ and 0, and hence is positive semi-definite. The matrix $\mathbf{B}$ is diagonally dominant with positive entries, since $a_{ii} = \sum_j a_{ij}^2$, $b_{ii} = 2a_{ii} - a_{ii}^2 = a_{ii} + \sum_{j\neq i} a_{ij}^2$, and $a_{ii} > 0$ implies

$$b_{ii} > a_{ii} > 0, \quad \text{and } b_{ii} > \sum_{j\neq i} a_{ij}^2 = \sum_{j\neq i} b_{ij}.$$

Hence, $\mathbf{B}$ is positive definite (Horn and Johnson, 1990). It follows that the Hessian matrix $\frac{\partial^2 \Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial \boldsymbol{\mu}\partial \boldsymbol{\mu}^\mathsf{T}}$ is positive semi-definite on $\mathcal{M}_n$ for PO-WR and multinomial sampling designs.

It remains to prove convexity for PO-WOR. First note that the function $\log\det(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$, by assumptions on $\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$, is differentiable and continuous on $\mathcal{M}_n$. It suffices, by continuity, to prove that the Hessian is positive semi-definite on the interior of $\mathcal{M}_n$. Consider therefore a point $\boldsymbol{\mu} \in \mathcal{M}_n$ such that $\mu_i < 1$ for all i. Let $\boldsymbol{u}_i = \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}\sqrt{1-\mu_i}/\sqrt{\mu_i}$ and $\mathbf{U}$ be the matrix with rows $\boldsymbol{u}_i^\mathsf{T}$. Also let $\mathbf{A} = \mathbf{U}(\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\mathbf{U}^\mathsf{T}$ and a_{ij} the elements of $\mathbf{A}$. Similar to above, we may now write

$$\frac{\partial^2\Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i\partial\mu_j} = \begin{cases} \mu_i^{-2}(1-\mu_i)^{-1}(2a_{ii} - a_{ii}^2) & \text{if } i = j, \\ (\mu_i\mu_j)^{-1}(1-\mu_i)^{-1/2}(1-\mu_j)^{-1/2}a_{ij}^2 & \text{if } i \neq j. \end{cases}$$

We recognise the Hessian matrix $\frac{\partial^2\Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}^\mathsf{T}}$ as the Hadamard product $\mathbf{M} \otimes \mathbf{B}$ of a rank-one matrix $\mathbf{M} = \mathbf{m}\mathbf{m}^\mathsf{T}$ with $\mathbf{m} = (\mu_1^{-1}(1-\mu_1)^{-1/2}, \ldots, \mu_N^{-1}(1-\mu_N)^{-1/2})^\mathsf{T}$, and a symmetric matrix $\mathbf{B}$ with entries

$$b_{ij} = \begin{cases} 2a_{ii} - a_{ii}^2 & \text{if } i = j, \\ a_{ij}^2 & \text{if } i \neq j. \end{cases}$$

The remainder of the proof follows in complete analogy with the proof for PO-WR and multinomial sampling designs.

A.7 Proof of Proposition 3

Proof of a) First note that the Hessian $\mathbf{H}_d(\boldsymbol{\theta})$ is positive semi-definite at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, since $\boldsymbol{\theta}_0$ is the global minimiser of $d(\boldsymbol{\theta})$. Hence, there exists a matrix $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}^\mathsf{T} = \mathbf{H}_d(\boldsymbol{\theta})$. $\mathbf{L}$ is non-zero since $\mathbf{H}_d(\boldsymbol{\theta})$, by assumption, is non-zero. The d-optimal sampling scheme $\boldsymbol{\mu}^*$ is defined as the minimiser of the function $\mathrm{tr}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{H}_d(\boldsymbol{\theta}_0)) = \mathrm{tr}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{L}\mathbf{L}^\mathsf{T})$, which by definition is equivalent to L-optimality with respect to a matrix $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}^\mathsf{T} = \mathbf{H}_d(\boldsymbol{\theta})$.

Proof of b) Assume that $\boldsymbol{\mu}^*$ is the minimser of $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ and let $d(\boldsymbol{\theta}) = \frac{1}{2}||\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)^\mathsf{T}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)||_2^2$ with Hessian matrix $\mathbf{H}_d(\boldsymbol{\theta}_0) = \mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)^\mathsf{T}$. According to Proposition 1, the Φ-optimal sampling scheme $\boldsymbol{\mu}^*$ must satisfy the optimality conditions (13) or (14a)–(14c) with

$$\begin{aligned} c_i &= ||\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)||_2^2 \\ &= \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0) \\ &= \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\mathbf{H}_d(\boldsymbol{\theta}_0)\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0). \end{aligned}$$

This is identical to the optimality conditions for the d-optimality criterion. Moreover, the d-optimality criterion is convex in $\boldsymbol{\mu}$ by Proposition 2a) and 3a), so $\boldsymbol{\mu}^*$ must be the global minimiser for the d-optimality criterion. Now, minimising $\mathrm{tr}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{H}_d(\boldsymbol{\theta}_0))$ is equivalent to minimising $\mathrm{tr}(k\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{H}_d(\boldsymbol{\theta}_0))$ for any constant $k > 0$, so $\boldsymbol{\mu}^*$ is also d-optimal with respect to the distance function $d(\boldsymbol{\theta}) = ||\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)^\mathsf{T}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)||_2^2$.

A.8 Proof of Proposition 4

The results follow from Proposition 3a) since the Hessian matrices of $d_{\mathrm{ER}}(\boldsymbol{\theta})$, $d_{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$, and d_{KL} are given by

$$\begin{aligned} \frac{\partial^2 d_{\mathrm{ER}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}} &= \frac{\partial^2\ell_0(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}} = \mathbf{H}(\boldsymbol{\theta}), \\ \frac{\partial^2 d_{\boldsymbol{\Sigma}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}} &= \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) = \boldsymbol{\Sigma}^{-1}, \text{ and} \\ \frac{\partial^2 d_{\mathrm{KL}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}} &= \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}}\sum_{i\in\mathcal{D}}\int_{\mathcal{Y}}\log\frac{f_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x}_i)}{f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}_i)}dF_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x}_i) \\ &= -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}}\sum_{i\in\mathcal{D}}\int_{\mathcal{Y}}\log f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}_i)dF_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x}_i) \\ &= -\sum_{i\in\mathcal{D}}\int_{\mathcal{Y}}\frac{\partial^2\log f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}_i)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}}dF_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x}_i) \\ &= \mathrm{E}_{\boldsymbol{y}\sim f_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x})}\left[-\sum_{i\in\mathcal{D}}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}}\log f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}_i)\right] \\ &= \mathrm{E}_{\boldsymbol{y}\sim f_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x})}[\mathbf{H}(\boldsymbol{\theta})] = \widetilde{\mathbf{H}}(\boldsymbol{\theta}). \end{aligned}$$

For the Hessian of the Kullback-Leibler distance we have used the Leibniz integral rule to change the order of integration and differentiation (cf. Kullback and Leibler, 1951).

A.9 Proof of Proposition 5

Consider a one-to-one differentiable mapping $\boldsymbol{g} : \boldsymbol{\theta} \mapsto \boldsymbol{\eta}$. Denote by $\ell_0^*(\boldsymbol{\eta}) = \sum_{i \in \mathcal{D}} \ell_i(\boldsymbol{g}^{-1}(\boldsymbol{\eta}))$ the induced empirical risk, with the minimiser $\boldsymbol{\eta}_0 = \boldsymbol{g}(\boldsymbol{\theta}_0)$. By the chain rule, the Hessian matrix of $\ell_0^*(\boldsymbol{\eta})$ at $\boldsymbol{\eta} = \boldsymbol{\eta}_0$ is given by

$$\begin{aligned}\mathbf{H}_{\ell_0^*}(\boldsymbol{\eta}_0) = \left.\frac{\partial^2 \ell_0^*(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^\mathsf{T}}\right|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} &= \mathbf{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{g}^{-1}(\boldsymbol{\eta}_0))^\mathsf{T} \mathbf{H}_{\ell_0}(\boldsymbol{g}^{-1}(\boldsymbol{\eta}_0)) \mathbf{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{g}^{-1}(\boldsymbol{\eta}_0)) \\ &= \mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-\mathsf{T}} \mathbf{H}(\boldsymbol{\theta}_0) \mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-1}.\end{aligned}$$

Here we have also used the fact that $\nabla_{\boldsymbol{\theta}} \ell_0(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \mathbf{0}$, by definition of $\boldsymbol{\theta}_0$ as the minimiser of $\ell_0(\boldsymbol{\theta})$.

Now assume that $\boldsymbol{\mu}^*$ and $\tilde{\boldsymbol{\mu}}^*$ are d_{ER}-optimal for $\boldsymbol{\theta}_0$ and $\boldsymbol{\eta}_0$, respectively. By the latter we mean that $\tilde{\boldsymbol{\mu}}^*$ minimises the expected distance of $\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}} = \boldsymbol{g}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}})$ from $\boldsymbol{\eta}_0 = \boldsymbol{g}(\boldsymbol{\theta}_0)$ with respect to the induced empirical risk distance $d_{\mathrm{ER}}^*(\boldsymbol{\eta}) = \ell_0^*(\boldsymbol{\eta}) - \ell_0^*(\boldsymbol{\eta}_0)$. By Proposition 4, $\boldsymbol{\mu}^*$ is L-optimal with respect to a matrix $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}^\mathsf{T} = \mathbf{H}(\boldsymbol{\theta}_0)$. Similarly, $\tilde{\boldsymbol{\mu}}^*$ is L-optimal with respect to a matrix $\widetilde{\mathbf{L}}$ such that

$$\widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\mathsf{T} = \mathbf{H}_{\ell_0^*}(\boldsymbol{\eta}_0) = \mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-\mathsf{T}} \mathbf{H}(\boldsymbol{\theta}_0) \mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-1} = \mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-\mathsf{T}} \mathbf{L}\mathbf{L}^\mathsf{T} \mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-1}. \tag{27}$$

Now, $\boldsymbol{\mu}^*$ is the minimiser of the function

$$\begin{aligned}\operatorname{tr}(\boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)\mathbf{L}\mathbf{L}^\mathsf{T}) &= \operatorname{tr}(\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-1}\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)\boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-\mathsf{T}}\mathbf{L}\mathbf{L}^\mathsf{T}) \\ &= \operatorname{tr}(\boldsymbol{\Gamma}_{\boldsymbol{g}}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-\mathsf{T}}\mathbf{L}\mathbf{L}^\mathsf{T}\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-1}) \\ &= \operatorname{tr}(\boldsymbol{\Gamma}_{\boldsymbol{g}}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)\widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\mathsf{T}).\end{aligned}$$

The first equality follows by inserting the identity matrix $\mathbf{I}_{p \times p} = \mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-1}\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0) = \mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-\mathsf{T}}$ twice, the second equality by (16) and the cyclic property of the trace, and the third equality by (27). But $\tilde{\boldsymbol{\mu}}^*$ is also a minimiser of $\operatorname{tr}(\boldsymbol{\Gamma}_{\boldsymbol{g}}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)\widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\mathsf{T})$. Since the L-optimality criterion is convex in $\boldsymbol{\mu}$, the optimum is unique and we must have $\boldsymbol{\mu}^* = \tilde{\boldsymbol{\mu}}^*$. Hence, the d_{ER}-optimality criterion is invariant under the re-parameterisation $\boldsymbol{g} : \boldsymbol{\theta} \mapsto \boldsymbol{\eta}$.

The results for d_S- and d_{KL}-optimality follow analogously.

B Supplementary Figures

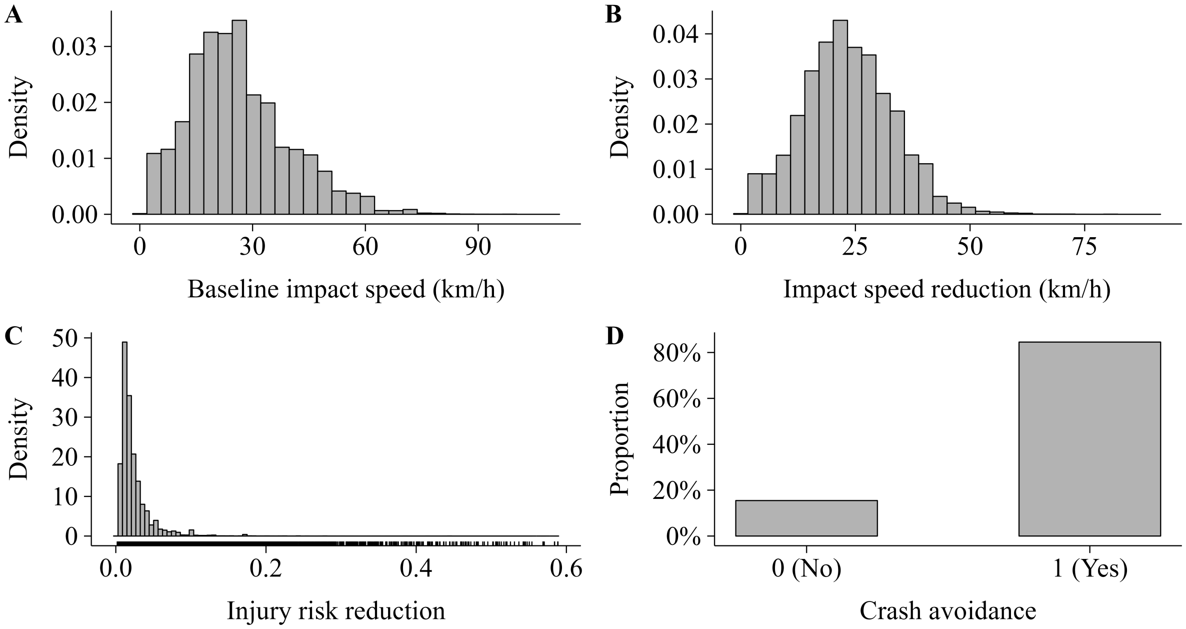

Figure S1. Characteristics of the vehicle safety assessment dataset considered in Section 6. **A**: Impact speed distribution under a baseline manual driving scenario. **B–D**: Distribution of the impact speed reduction, injury risk reduction, and crash avoidance rate, with an automatic emergency system compared to the baseline manual driving scenario.

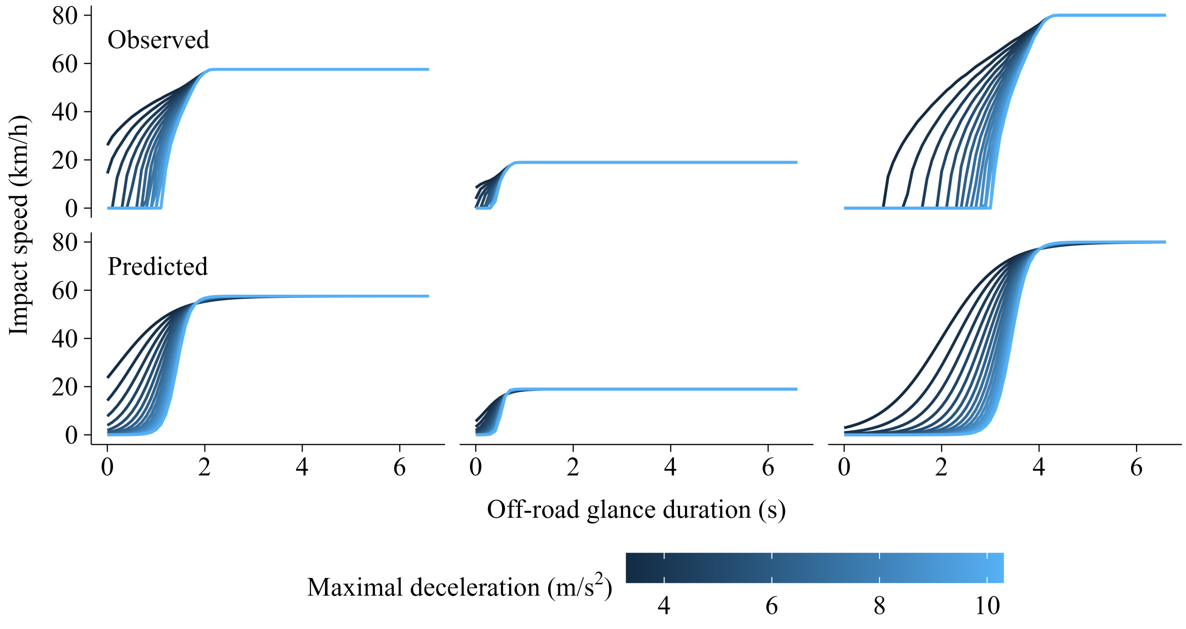

Figure S2. Impact speed response surface as a function of off-road glance duration and maximal deceleration during braking for counterfactual variations of three reconstructed rear-end crashes. **Top panel**: Observed impact speed. **Bottom panel**: Predicted impact speed using the quasi-binomial logistic regression model (21). The response values have been mapped from the model range $[0, 1]$ to the original range $[0, y_{\max,k}]$, where $y_{\max,k}$ is the maximal possible impact speed for the variations generated from case k, $k = 1, \dots, 44$.