Active sampling: A machine-learning-assisted framework for finite population inference with optimal subsamples

Henrik Imberg[1], Xiaomi Yang[2], Carol Flannagan[2,3], Jonas Bärgman[2]

[1]Department of Mathematical Sciences,
Chalmers University of Technology and University of Gothenburg

[2]Division of Vehicle Safety, Chalmers University of Technology

[3]University of Michigan Transportation Research Institute

Abstract

Data subsampling has become widely recognized as a tool to overcome computational and economic bottlenecks in analyzing massive datasets and measurement-constrained experiments. However, traditional subsampling methods often suffer from the lack of information available at the design stage. We propose an active sampling strategy that iterates between estimation and data collection with optimal subsamples, guided by machine learning predictions on yet unseen data. The method is illustrated on virtual simulation-based safety assessment of advanced driver assistance systems. Substantial performance improvements were observed compared to traditional sampling methods.

Keywords: active learning, adaptive importance sampling, inverse probability weighting, survey sampling, optimal design

1 Introduction

Enabled by advances of modern technology, data are currently being generated at greater volumes than ever before. This poses major challenges for statistical methods and data analysis procedures. Sometimes the sheer amount of data is too large to be stored, processed and analyzed in reasonable time with available resources (Ma and Sun, 2015). In other

cases, observing complete data may be expensive and hence affordable only for a subset of a large initial dataset, a situation known as a measurement-constrained experiment (Wang et al., 2017; Meng et al., 2021). In either case, analysis of complete data is prohibited by computational, practical, economic or even ethical cost constraints. In such circumstances, researchers often resort to subsampling.

The origin of modern data subsampling methods dates back to the early 1900's, where probability sampling and estimation through inverse probability weighting were introduced as tools for inference about finite population characteristics. Pioneered by work of Neyman (1938), Hansen and Hurwitz (1943), Horvitz and Thompson (1952) and others, this evolved to the subfield of statistics today known as survey sampling (Tille, 2020). With the advancement of big data, sampling methods have gained renewed attention beyond the classical problems in finite population inference. Recent developments include regression modeling with massive datasets (Ma et al., 2015; Wang et al., 2018; Dai et al., 2022), large scale Bayesian inference and Markov Chain Monte Carlo methods (Magnusson et al., 2019; Quiroz et al., 2021), and active learning (Bach, 2007; Beygelzimer et al., 2009; Ganti and Gray, 2012; Imberg et al., 2020). Areas of application include pattern recognition and image classification (Kossen et al., 2021; Farquhar et al., 2021), analysis of naturalistic driving studies (Mousa et al., 2019; Imberg et al., 2022), and scenario generation for virtual safety assessment of advanced driver assistance systems and autonomous driving systems (de Gelder and Paardekooper, 2017; Akagi et al., 2019; Wang et al., 2021; Zhang et al., 2021), to mention a few.

The crucial question for any subsampling method is how subsamples should be selected for optimal performance. The dilemma of optimal design is that actual optimal design requires prior knowledge of the phenomenon which the experiment is intended to study. However, if such information were available, the experiment would not need to be run. Consequently, the "optimal design" inevitably depends on assumptions that only can be tested after data has been collected. Active learning offers a promising solution to this problem. An active learner is a machine learning algorithm that itself chooses the data from which it learns, typically from a large set of potential training examples. This is done in an iterative manner by selecting a small batch of new training examples, retrieving the corresponding data, and updating the predictions of outcomes on yet unseen data. Guided by these predictions, new instances can be selected in an optimal manner and the process repeated until a sufficient amount of data has been collected (Settles, 2012). Although initially developed for prediction research, active learning procedures may be utilized to derive machine-learning-assisted optimal sampling schemes in general subsampling problems.

We consider the problem of estimating a simple finite population characteristic, such as a total, mean, or ratio. This classical problem has received much attention in the survey sampling literature (cf. Neyman, 1938; Hansen and Hurwitz, 1943; Horvitz and Thompson, 1952; Tille, 2020). However, the possibilities offered by machine learning in this context have not been fully explored. Traditional subsampling methods often suffer from the lack of information available at the design stage, rendering optimal sampling practically unfeasible. To address this issue, we propose an active sampling strategy where such information is acquired gradually during the data collection process, with optimal subsamples guided by machine learning predictions on yet unseen data.

Outline The structure of this paper is as follows: We start by presenting a motivating example in crash-causation-based scenario generation for virtual vehicle safety assessment in Section 2. The mathematical framework is introduced in Section 3, where we also review traditional methods for finite population inference through unequal probability sampling and inverse probability weighting. The main contribution of this work is presented in Section 4, where we introduce active sampling to increase the efficiency in estimating a finite population characteristic, such as a total, mean, or ratio. Additional theoretical results and proofs are provided in Appendix A. Empirical results in crash-causation-based scenario generation for virtual vehicle safety assessment is presented in Section 5.

2 Motivating example

Traffic safety is a substantial problem worldwide (World Health Organization, 2018). Safety systems have been developed to improve traffic safety and have shown the potential to avoid or mitigate crashes. However, when developing both advanced driver assistance systems (ADAS) and autonomous driving systems (ADS), there is a need to assess the impact on safety of the systems before they are on the market. One way to do that is by running virtual simulations comparing the outcome of simulations both with and without a specific system (Anderson et al., 2013; Sander, 2018; Seyedi et al., 2021). Such simulations are often called counterfactual simulations, as they assess what could have happened if the system were on the road.

A prerequisite of such simulations is to have baseline crashes — a set of pre-crash kinematics of crashes, describing how the involved road users move prior to the crash, onto which the system can be (virtually) applied. One source of baseline pre-crash kinematics is the generation of crashes by applying crash-causation models on the pre-crash kinematics

from reconstructed real-world crashes. For example, Bärgman et al. (2022) validated such a crash-causation model for the generation of rear-end crashes.

We consider scenario generation based on a glance-and-deceleration crash-causation model where a driver's off-road glance behavior and braking profile are represented by discrete (empirical) probability distributions. The outcome of the simulations is a distribution of impact speeds of all the crashes generated by all combinations of the eyes-off-road glace duration and the maximum deceleration during braking. Here "all combinations" is the problem. Complete enumeration becomes practically unfeasible in high-dimensional (many parameters varied) or high-resolution (many levels per parameter) settings, and subsampling inevitable. There is in scenario generation a need for efficient, flexible and data-driven methods for sample selection.

3 Finite population sampling

We first introduce the mathematical framework and notation in Section 3.1, presented in the context of the crash-causation-based scenario generation application outlined in Section 2. Next, some basics of unequal probability sampling and estimation through inverse probability weighting are reviewed in Section 3.2.

3.1 Mathematical framework and setup

Assume we are given a finite population or dataset $\mathcal{D}$ with N instances or elements $i = 1, \ldots, N$. Associated with each element i in $\mathcal{D}$ is a collection of variables $(p_i, r_i, \boldsymbol{y}_i^T, \boldsymbol{z}_i^T)$, where p_i is a prior weight or observation weight associated with element i; r_i a binary inclusion/exclusion indicator variable, taking the value 1 if element i is relevant for the scientific question of interest and should be included in the analyses, and 0 otherwise; $\boldsymbol{y}_i$ a vector of outcomes or response variables, and $\boldsymbol{z}_i$ a vector of design variables and auxiliary variables. We use scalar notation y_i to denote a single element of the response vector $\boldsymbol{y}_i$. Vectors are assumed to be column vectors unless otherwise stated.

In the context of crash-causation-based scenario generation, the dataset $\mathcal{D}$ represents a collection of N potential simulation scenarios of interest. The observation weights p_i are included to account for the probabilities of the different scenarios occurring in real life. The response variables $\boldsymbol{y}_i$ are outcomes of the simulation, including, e.g., whether a crash occurred or not, impact speed if there was a crash, and impact speed reduction with an advanced driver assistance system (ADAS) compared to some baseline driving scenario. The auxiliary variables $\boldsymbol{z}_i$ contain scenario information, such as simulation settings and

parameters that are under the control of the investigator, and any additional information that is available without running the actual simulation. Since the aim in the current application is to evaluate the safety benefit of an ADAS, and it is assumed that the number of crashes created by the ADAS itself is zero or very small, we are only interested in simulations that produce a crash under some baseline driving scenarios. This is captured by the binary relevance indicator variable r_i, taking the value 1 if there is a crash in the baseline scenario, and 0 otherwise. Inference will be restricted to the subset of simulation scenarios for which $r_i = 1$. The variables p_i and $\boldsymbol{z_i}$ are available *a priori* for all members $i \in \mathcal{D}$, while the variables r_i and $\boldsymbol{y}_i$ only can be observed by running the corresponding virtual simulation.

The scope of inference in this paper is on classical survey sampling tasks, e.g., estimating a total or functions of totals. Specifically, we will consider following characteristics of the dataset $\mathcal{D}$:

i) a total $t_y = \sum_{i \in \mathcal{D}} p_i r_i y_i$,

ii) the mean among relevant instances, given by the ratio $t_y/t_r, t_r = \sum_{i \in \mathcal{D}} p_i r_i$,

iii) and, more generally, a population characteristic θ given by $\theta = h(\boldsymbol{t_y})$ for some differentiable function $h : \mathbb{R}^d \to \mathbb{R}$ and d-dimensional vector of totals $\boldsymbol{t_y} = \sum_{i \in \mathcal{D}} p_i r_i \boldsymbol{y}_i$.

We note that the observation weights p_i and relevance indicators r_i may not be present or needed in all applications. In such case, they can simply be ignored or set equal to one in the formulas above and presentation that follows. We consider the total t_y and mean t_y/t_r as estimation of means and totals are standard tasks in the survey sampling and Monte Carlo literature (Fishman, 1996; Tille, 2020), and also often encountered in the scenario generation context (Wang et al., 2021). The main interest in our application lies in the mean t_y/t_r, as this is what we use to calculate, e.g., mean impact speed reduction and crash avoidance rate with an advanced driver assistance system compared to some baseline driving scenario, when restricted to the relevant set of crashes. We also consider smooth functions of totals to provide some general results. Note that i) and ii) are included in the third class of statistics. To see this, take $\boldsymbol{y}_i = y_i$ and $h(u) = u$ to obtain i), and $\boldsymbol{y}_i = (y_i, 1)^T$ and $h(u_1, u_2) = u_1/u_2$ to obtain ii). This class of statistics also includes, e.g., ratios, correlation coefficients, and population variances.

3.2 Unequal probability sampling and estimation

In our crash-causation based scenario generation application, running all N simulations of interest to observe complete data $(p_i, r_i, \boldsymbol{y}_i^T, \boldsymbol{z}_i^T)$ for all members $i \in \mathcal{D}$ is too high dimensional (many simulation parameters varied) or high-resolution (many levels per parameter) to be feasible in practice. Also, even if complete enumeration were feasible, it may not be efficient. It is not implausible to assume a good estimate can be obtained with much lower computational demand. In other applications, observing complete data may be hampered due to other costs, including, e.g., monetary, computational, or ethical costs. Thus, we assume that observing complete data is costly and only affordable for a subset $\mathcal{S} \subset \mathcal{D}$ of size n.

We consider the case where the subset $\mathcal{S}$ for which complete data will be observed is selected using unequal probability sampling. To allow for sampling with replacement, we let S_i be the random variable representing the number of times an element $i \in \mathcal{D}$ is selected by the sampling mechanism, and denote by $\mu_i := \mathrm{E}[S_i]$ the corresponding mean. Hence, $\mathcal{S}$ is the random set given by $\mathcal{S} = \{i \in \mathcal{D} : S_i > 0\}$. Some common unequal probability sampling designs include multinomial sampling, Poisson sampling (with or without replacement), and adjusted Poisson sampling (Tillé, 2006).

To account for unequal probabilities of selection, estimation may be performed by sample weighting techniques using, e.g., the Hansen-Hurwitz estimator (Hansen and Hurwitz, 1943)

$$\hat{t}_y = \sum_{i \in \mathcal{S}} S_i w_i p_i r_i y_i, \quad w_i = 1/\mu_i \tag{1}$$

This is an unbiased estimator of the total t_y, provided that $\mu_i > 0$ for all $i \in \mathcal{D}$. To see this, simply note that $\hat{t}_y = \sum_{i \in \mathcal{S}} S_i w_i p_i r_i y_i = \sum_{i \in \mathcal{D}} S_i w_i p_i r_i y_i$, and $\mathrm{E}[S_i] = \mu_i = w_i^{-1}$. Similarly, an estimator for the mean among relevant instances is given by $\hat{t}_y/\hat{t}_r$ with $\hat{t}_r = \sum_{i \in \mathcal{S}} S_i w_i p_i r_i$, and estimator for a characteristic $\theta := h(\boldsymbol{t}_{\boldsymbol{y}})$ by $\hat{\theta} = h(\hat{\boldsymbol{t}}_{\boldsymbol{y}})$, with $\hat{\boldsymbol{t}}_{\boldsymbol{y}} = \sum_{i \in \mathcal{S}} S_i w_i p_i r_i \boldsymbol{y}_i$. Note that $\hat{t}_y$ and $\hat{t}_y/\hat{t}_r$ are special cases of the latter more general class of estimators. To see this, take $\boldsymbol{y}_i = y_i$ and $h(u) = u$ to obtain $\hat{\theta} = h(\hat{t}_y) = \hat{t}_y$, and $\boldsymbol{y}_i = (y_i, 1)^T$ and $h(u_1, u_2) = u_1/u_2$ to obtain $\hat{\theta} = h(\hat{\boldsymbol{t}}_{\boldsymbol{y}}) = \hat{t}_y/\hat{t}_r$. When $p_i = r_i = 1$ for all i, the ratio $\hat{t}_y/\hat{t}_r$ is coincides with the Hájek estimator $\left(\sum_{i \in \mathcal{S}} S_i w_i\right)^{-1} \sum_{i \in \mathcal{S}} S_i w_i y_i$ of the mean t_y/N. In this case, the alternative estimator $\hat{t}_y/N$ may also be used.

4 Active sampling

Consider for the moment estimation of a total t_y, and assume that $r_i = p_i = 1$ and $y_i > 0$ for all elements $i \in \mathcal{D}$. It then is a widely known fact that the optimal sampling scheme

for the estimator (1) is to sample with probability proportional to y_i (cf. Tillé, 2006). In practice, however, y_i are inaccessible at the design stage and the optimal sampling scheme is therefore unknown. Inspired by active learning (Settles, 2012), we propose in Section 4.1 an active sampling strategy that addresses this issue by iterating between parameter estimation and data collection with optimal subsamples guided by machine learning predictions on yet unseen data. Optimal sampling schemes for estimating simple finite population characteristics, such as totals and functions of totals, are presented in Section 4.2. To simplify the presentation we will restrict ourselves to multinomial sampling designs, but note that the procedure may be easily adapted to other unequal probability sampling designs.

4.1 Active sampling algorithm, estimation, and inference

The active sampling method is summarized in Algorithm 1. The algorithm proceeds in K iterations $k = 1, \ldots, K$ and chooses, in each iteration, n_k new instances at random (possibly with replacement) from $\mathcal{D}$. Once a new batch of instances has been selected we observe the corresponding data $(r_i, \boldsymbol{y}_i^T)$ and update our estimates of the characteristics of interest. The process continues until a pre-specified maximal number of iterations K is reached, or the target characteristic is estimated with sufficient precision. Following the notation in the previous section, we let S_{ki} be the random variable representing the number of times an element $i \in \mathcal{D}$ is selected in iteration k. We let $\boldsymbol{S}_k = (S_{k1}, \ldots, S_{kN})$, and denote by $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{kN}) = \mathrm{E}[\boldsymbol{S}_k | \boldsymbol{S}_1, \ldots, \boldsymbol{S}_{k-1}]$ the conditional expectation of $\boldsymbol{S}_k$ given the previous selections $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_{k-1}$. We assume that $\boldsymbol{S}_k | \boldsymbol{S}_1, \ldots \boldsymbol{S}_{k-1} \sim \text{Multinomial}(n_k, \boldsymbol{\pi}_k)$, so that $\boldsymbol{\mu}_k = n_k \boldsymbol{\pi}_k$. We will refer to the probability vector $\boldsymbol{\pi}_k = (\pi_{k1}, \ldots, \pi_{kN})$ as the sampling scheme in iteration k. As usual, we require that $\pi_{ki} > 0$ for all k and all $i \in \mathcal{D}$, although technically one could allow for $\pi_{ki} = 0$ whenever r_i is known to be zero.

A key component of the active sampling algorithm is the inclusion of an auxiliary model $f(r_i, \boldsymbol{y}_i^T | \boldsymbol{z}_i^T)$ for the joint distribution of the unobserved data $(r_i, \boldsymbol{y}_i^T)$ given auxiliary variables $\boldsymbol{z}_i$. At this stage, any prediction model or machine learning algorithm can be used. By gathering data in a sequential manner, we may iteratively update our predictions on yet unseen data. Doing so, we are able to learn from past observations how to sample in an optimal way in future iterations (Subroutine 1). Further details on optimal sampling schemes for estimating a total or function of totals are provided in Section 4.2. We next describe the process for inference in active sampling and the asymptotic properties of active sampling estimators in Sections 4.1.1–4.1.3 below.

Algorithm 1 Active sampling

Input: Sampling frame $\mathcal{D}$, target characteristic $\theta = h(\boldsymbol{t}_{\boldsymbol{y}})$ (to be estimated), precision target δ, maximal number of iterations K, batch sizes $\{n_k\}_{k=1}^{K}$.
Initialization: Let $m_0 = 0$, $\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(0)} = \mathbf{0}$, $\mathcal{L}_0 = \varnothing$.

1: **for** k = 1, 2, . . . , K **do**
2: **Learning** (if $k > 1$): Train prediction model $f(r_i, \boldsymbol{y}_i^T | \boldsymbol{z}_i^T)$ on the labeled dataset $\{(s_{ji}, \mu_{ji}, p_i, r_i, \boldsymbol{y}_i^T, \boldsymbol{z}_i^T)\}_{i \in \mathcal{L}_j, j=1,\ldots,k-1}$.
3: **if** k > 1 and **Learning** step was successful† **then**
4: **Optimization**: Calculate sampling scheme $\boldsymbol{\pi}_k$ according to Subroutine 1.
5: **else**
6: **Fallback**: Set $\pi_{ki} = \frac{p_i}{\sum_{j \in \mathcal{D}} p_j}$ for all $i \in \mathcal{D}$.
7: **end if**
8: **Sampling**: Draw vector $\boldsymbol{s}_k = (s_{k1}, \ldots, s_{kN}) \sim \text{Multinomial}(n_k, \boldsymbol{\pi}_k)$.
9: **Labeling**: Retrieve data $(r_i, \boldsymbol{y}_i^T)$ for selected instance(s) $i \in \mathcal{L}_k := \{i \in \mathcal{D} : s_{ki} > 0\}$.
10: **Estimation**: Let

$$\hat{\boldsymbol{t}}_{\boldsymbol{y},k} = \sum_{i \in \mathcal{L}_k} s_{ki} w_{ki} p_i r_i \boldsymbol{y}_i, \quad w_{ki} = 1/\mu_{ki},$$

$$m_k = m_{k-1} + n_k, \quad \hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)} = \frac{1}{m_k} \left(m_{k-1} \hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k-1)} + n_k \hat{\boldsymbol{t}}_{\boldsymbol{y},k} \right), \quad \hat{\theta}^{(k)} = h(\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)}).$$

11: Estimate the variance of $\hat{\theta}^{(k)}$ according to (3).
12: **if** $\widehat{\text{Var}}(\hat{\theta}^{(k)}) < \delta$ **then**
13: **Termination**: Stop algorithm. Continue to 16.
14: **end if**
15: **end for**
16: **Output**: Estimate $\hat{\theta}^{(k)}$ and labeled dataset $\{(s_{ji}, \mu_{ji}, p_i, r_i, \boldsymbol{y}_i^T, \boldsymbol{z}_i^T)\}_{i \in \mathcal{L}_j, j=1,\ldots,k}$ of records with complete data $(p_i, r_i, \boldsymbol{y}_i^T, \boldsymbol{z}_i^T)$.

†By a successful learning step we mean that the prediction model could be fitted and reliable predictions obtained, as assessed by some measure of generalization error (e.g., accuracy or coefficient of determination on hold-out data).

4.1.1 Estimating a finite population characteristic

As more data becomes available, we iteratively update our estimates of the population characteristics of interest by similar means as in (1). Specifically, we first construct an

Subroutine 1 Optimal sampling schemes

Input: Sampling frame $\mathcal{D}$, target characteristic $\theta := h(\boldsymbol{t_y})$ (to be estimated), estimate $\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k-1)}$ of $\boldsymbol{t_y}$, predictions $\hat{r}_i, \hat{\boldsymbol{y}}_i$ of r_i and $\boldsymbol{y}_i | r_i = 1$ derived from the prediction model $f(r_i, \boldsymbol{y}_i^T | \boldsymbol{z}_i^T)$, and estimates $\hat{\boldsymbol{\Sigma}}_i$ of the residual covariance matrices $\boldsymbol{\Sigma}_i := \mathbf{Cov}(\boldsymbol{Y}_i - \hat{\boldsymbol{y}}_i | r_i = 1)$, with the unknown values of $(r_i, \boldsymbol{y}_i^T)$ treated as random variables $(R_i, \boldsymbol{Y}_i^T)$ distributed according to $f(r_i, \boldsymbol{y}_i^T | \boldsymbol{z}_i^T)$.

if $\nabla h(\boldsymbol{u})\big|_{\boldsymbol{u}=\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k-1)}} = \mathbf{0}$ or undefined **then**

Set

$$\pi_{ki} = \frac{p_i}{\sum_{j \in \mathcal{D}} p_j} \quad \text{for all } i \in \mathcal{D}.$$

else

Calculate sampling probabilities according to:

$$\pi_{ki} = \frac{p_i \sqrt{v_i}}{\sum_{j \in \mathcal{D}} p_j \sqrt{v_j}} \quad \text{for all } i \in \mathcal{D}.$$

with v_i given as in Method 1 or 2 below.[†]

Method 1 (Naive):

$$v_i = \hat{r}_i^2 \big| \nabla h(\boldsymbol{u})^T \hat{\boldsymbol{y}}_i \big|^2_{\boldsymbol{u}=\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k-1)}}.$$

Method 2 (Minimize anticipated variance):

$$v_i = \hat{r}_i [(\nabla h(\boldsymbol{u})^T \hat{\boldsymbol{y}}_i)^2 + \nabla h(\boldsymbol{u})^T \hat{\boldsymbol{\Sigma}}_i \nabla h(\boldsymbol{u})]\big|_{\boldsymbol{u}=\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k-1)}}.$$

end if

Output: Sampling scheme $\boldsymbol{\pi}_k = (\pi_{k1}, \ldots, \pi_{kN})$.

[†]See Section 4.2 for further details. Simplified formulas of sampling schemes for estimating a total t_y or mean t_y/t_r are presented in Corollary 1 and 2.

unbiased estimator $\hat{\boldsymbol{t}}_{\boldsymbol{y},k}$ of a vector of totals $\boldsymbol{t_y}$ according to

$$\hat{\boldsymbol{t}}_{\boldsymbol{y},k} = \sum_{i \in \mathcal{D}} S_{ki} w_{ki} p_i r_i \boldsymbol{y}_i, \quad w_{ki} = 1/\mu_{ki}.$$

Given one such estimator from each of the preceding iterations, we construct a pooled estimator $\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)}$ as

$$\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)} = \frac{1}{m_k} \sum_{j=1}^{k} n_j \hat{\boldsymbol{t}}_{\boldsymbol{y},j}, \tag{2}$$

where $m_k = n_1 + \ldots + n_k$ is the total sample size after k iterations of the active sampling algorithm. An alternative recursive formulation is provided in Algorithm 1. An estimator of a characteristic θ defined by $\theta = h(\boldsymbol{t}_{\boldsymbol{y}}^{(k)})$ for some function $h : \mathbb{R}^d \to \mathbb{R}$ is then obtained by $\hat{\theta}^{(k)} = h(\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)})$.

4.1.2 Variance estimation

To estimate the variance of our estimator $\hat{\theta}^{(k)}$, we first need an estimator of the covariance matrix $\boldsymbol{\Psi}^{(k)} = \mathbf{Cov}(\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)})$ of $\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)}$. Three such estimators are presented below. A theoretical justification is provided by Proposition S1 and Corollary S1 in Appendix A.1.

Method 1 (Classical method): First, we may proceed in analogy to (2) and use the pooled estimator

$$\hat{\boldsymbol{\Psi}}_1^{(k)} = m_k^{-2} \sum_{j=1}^{k} n_j^2 \hat{\boldsymbol{\Phi}}_j,$$

where $\hat{\boldsymbol{\Phi}}_j$ are (any) unbiased estimators of the conditional covariance matrices $\boldsymbol{\Phi}_j = \mathbf{Cov}(\hat{\boldsymbol{t}}_{\boldsymbol{y},j} | \boldsymbol{S}_1, \ldots, \boldsymbol{S}_{j-1})$. Each of the covariance matrices $\boldsymbol{\Phi}_j$ may be estimated using standard survey sampling techniques. For instance, under the multinomial design we may use Sen-Yates-Grundy estimator (Sen, 1953; Yates and Grundy, 1953) for $\boldsymbol{\Phi}_j$, i.e.,

$$\hat{\boldsymbol{\Phi}}_j = \frac{n_j}{n_j - 1} \sum_{i \in \mathcal{D}} S_{ji} \left(\frac{p_i r_i \boldsymbol{y}_i}{\mu_{ji}} - \frac{\hat{\boldsymbol{t}}_{\boldsymbol{y},j}}{n_j} \right) \left(\frac{p_i r_i \boldsymbol{y}_i}{\mu_{ji}} - \frac{\hat{\boldsymbol{t}}_{\boldsymbol{y},j}}{n_j} \right)^T, \quad \mu_{ji} = n_j \pi_{ji},$$

provided that $n_j \geq 2$. See, e.g., Tillé (2006, Chapter 5) for variance estimators under other unequal probability sampling designs.

Method 2 (Martingale method): Alternatively, we may use the squared variation of the estimates $\hat{\boldsymbol{t}}_{\boldsymbol{y},j}$ to estimate $\boldsymbol{\Psi}^{(k)}$ by

$$\hat{\boldsymbol{\Psi}}_2^{(k)} = m_k^{-2} \sum_{j=1}^{k} n_j^2 \left(\hat{\boldsymbol{t}}_{\boldsymbol{y},j} - \hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)} \right) \left(\hat{\boldsymbol{t}}_{\boldsymbol{y},j} - \hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)} \right)^T.$$

This method works also when $n_j = 1$, but generally requires the number of iterations k to be large.

Method 3 (Bootstrap method): Finally, variance estimation may be conducted by non-parametric bootstrap (Efron, 1979; Davison and Hinkley, 1997). If subsampling is done with replacement, importance weighted bootstrap should be used to account for possible differences in the number of selections per observation. Specifically, the bootstrap sample size should be equal to the total sample size $\sum_{j=1}^{k}\sum_{i\in\mathcal{D}} s_{ji}$, and selection probabilities proportional to the number of selections s_{ji}. One way to achieve this with ordinary bootstrap software is to create an extended dataset with on record for each of the s_{ji} selections, and perform ordinary non-parametric bootstrap on the extended dataset. An estimate of the covariance matrix of $\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)}$ is then obtained by

$$\hat{\boldsymbol{\Psi}}_3^{(k)} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\tilde{\boldsymbol{t}}_{\boldsymbol{y},b}^{(k)} - \bar{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)}\right)\left(\hat{\boldsymbol{t}}_{\boldsymbol{y},b}^{(k)} - \bar{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)}\right)^T,$$

where $\bar{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)} = \frac{1}{B}\sum_{b=1}^{B}\tilde{\boldsymbol{t}}_{\boldsymbol{y},b}^{(k)}$ is the mean of B bootstrap estimates $\tilde{\boldsymbol{t}}_{\boldsymbol{y},b}^{(k)}$ of $\boldsymbol{t}_{\boldsymbol{y}}$.

Given an estimate $\hat{\boldsymbol{\Psi}}^{(k)}$ of the covariance matrix of $\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)}$, the variance of $\hat{\theta}^{(k)} = h(\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)})$ may be estimated using the Delta method as

$$\widehat{\mathrm{Var}}(\hat{\theta}^{(k)}) = \nabla h(\boldsymbol{u})^T \hat{\boldsymbol{\Psi}}^{(k)} \nabla h(\boldsymbol{u})\big|_{\boldsymbol{u}=\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)}}, \tag{3}$$

provided that $\nabla h(\boldsymbol{u})\big|_{\boldsymbol{u}=\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)}} \neq \boldsymbol{0}$ (cf. Sen and Singer, 1993).

4.1.3 Interval estimation and asymptotic properties

Confidence intervals can be calculated using the classical large sample formula

$$\hat{\theta}^{(k)} \pm z_{\alpha/2} \times \mathrm{SE}_{\hat{\theta}^{(k)}} \tag{4}$$

where $\hat{\theta}^{(k)}$ is the estimate of the characteristic θ, $\mathrm{SE}_{\hat{\theta}^{(k)}} = \sqrt{\widehat{\mathrm{Var}}(\hat{\theta}^{(k)})}$ the corresponding standard error, and $z_{\alpha/2}$ the $\alpha/2$-quantile of a standard normal distribution. Under the assumptions of Proposition S1 and Corollary S1 in Appendix A.1, such a confidence interval has approximately $100 \times (1-\alpha)\%$ coverage of the true population characteristic θ, under repeated subsampling from $\mathcal{D}$, in large enough samples.

Using the martingale central limit theorem of Brown (1971), we show that under general regularity conditions our active sampling estimators are consistent and asymptotically normally distributed, for fixed N and bounded batch sizes $n_k < N$, as the number of iterations k tends to infinity (Proposition S1 and Corollary S1 in Appendix A.1). Our assumptions include:

i) standard moment conditions on the sample selection variables and selection probabilities, which essentially requires the sampling probabilities to be properly bounded away from zero,

ii) standard conditions on the total variance $\mathrm{Var}\left(\sum_{j=1}^{k} n_j \hat{\boldsymbol{t}}_{\boldsymbol{y},j}\right)$, which is assumed to tend to infinity, and

iii) that the sum of conditional variances $\sum_{j=1}^{k} \mathrm{Var}\left(n_j \hat{\boldsymbol{t}}_{\boldsymbol{y},j} | \boldsymbol{S}_1, \ldots, \boldsymbol{S}_{j-1}\right)$ behaves asymptotically like the total variance $\mathrm{Var}\left(\sum_{j=1}^{k} n_j \hat{\boldsymbol{t}}_{\boldsymbol{y},j}\right)$, i.e., the dependencies between the iterations of the active sampling algorithm should be asymptotically negligible. This ensures that the statistical properties of an active sampling estimator can be deduced from a single execution of the algorithm.

Empirical justification for these assumptions is provided in Section 5.

4.2 Optimal sampling schemes

We now turn to the question of sample selection and optimal choice of sampling scheme. These are crucial components in any subsampling method to obtain accurate estimates and reduce sample size requirements or, as in our application, computational complexity. Theoretically optimal sampling schemes for estimating functions of totals are presented in Section 4.2.1. However, these require knowledge about the full data $\{(r_i, \boldsymbol{y}_i^T)\}_{i \in \mathcal{D}}$ to be evaluated, and hence are of limited practical use. In Section 4.2.2 we present practically useful sampling schemes to minimize the expected variance of an estimator under an assisting auxiliary model for the unknowns. For proofs we refer to Appendix A.2.

4.2.1 Theoretical optimality

To derive an optimal sampling scheme, we need to define an objective function to be minimized. Since we consider estimation of a scalar characteristic $t_y, t_y/t_r$ or $\theta = h(\boldsymbol{t}_{\boldsymbol{y}})$, we aim to minimize the variance of the corresponding active sampling estimator as a function of the sampling schemes $\boldsymbol{\pi}_k$. This is, however, complicated by two facts:

i) explicit finite-sample variance formulas are generally not available for characteristics defined by non-linear functions of totals, and

ii) active sampling introduces complex dependencies between the different iterations of the algorithm.

Hence, finding a globally optimal sampling strategy for a finite number of iterations may not be feasible. We therefore resort to asymptotic arguments. First, to address i) we consider the approximate variance $\mathrm{AV}(\hat{\theta}^{(k)})$ of our estimator $\hat{\theta}^{(k)} = h(\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)})$, which by the Delta method is given by

$$\mathrm{AV}(\hat{\theta}^{(k)}) = \nabla h(\boldsymbol{u})^T \boldsymbol{\Psi}^{(k)} \nabla h(\boldsymbol{u})\big|_{\boldsymbol{u}=\boldsymbol{t}_{\boldsymbol{y}}},$$

provided that $h(\boldsymbol{u})\big|_{\boldsymbol{u}=\boldsymbol{t}_{\boldsymbol{y}}} \neq \mathbf{0}$, where $\boldsymbol{\Psi}^{(k)} = \mathbf{Cov}(\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)})$ (cf. Sen and Singer, 1993). If $\boldsymbol{\Psi}^{(k)}$ can be expressed as a simple function of the sampling schemes $\boldsymbol{\pi}_k$, then so can $\mathrm{AV}(\hat{\theta}^{(k)})$. To address ii), we note that the total covariance matrix $\boldsymbol{\Psi}^{(k)}$ is tightly connected to the sum of the conditional covariances $\boldsymbol{\Phi}_j := \mathbf{Cov}(\hat{\boldsymbol{t}}_{\boldsymbol{y},j}|\boldsymbol{S}_1, \ldots, \boldsymbol{S}_{j-1})$, $j \leq k$. Indeed, if we replace $\boldsymbol{\Psi}^{(k)}$ by the weighted sum $m_k^{-2} \sum_{j=1}^k n_j^2 \boldsymbol{\Phi}_j$, we obtain under the assumptions of Corollary S1 in Appendix A.1 another asymptotically valid expression for the variance of $\hat{\theta}^{(k)}$:

$$\widetilde{\mathrm{AV}}(\hat{\theta}^{(k)}) = m_k^{-2} \sum_{j=1}^{k} n_j^2 \nabla h(\boldsymbol{u})^T \boldsymbol{\Phi}_j \nabla h(\boldsymbol{u})\big|_{\boldsymbol{u}=\boldsymbol{t}_{\boldsymbol{y}}},$$

where $m_k = n_1 + \ldots + n_k$ is the total sample size after k iterations of the active sampling algorithm. That is, it holds under the assumptions of Corollary S1 that the limit of $m_k \widetilde{\mathrm{AV}}(\hat{\theta}^{(k)})$ is equal to the asymptotic variance of $\sqrt{m_k}(\hat{\theta}^{(k)} - \hat{\theta})$ with probability 1, as the number of iterations k tends to infinity. Hence, to minimize the variance of $\hat{\theta}^{(k)}$ we should minimize the approximate conditional variance $\nabla h(\boldsymbol{u})^T \boldsymbol{\Phi}_j \nabla h(\boldsymbol{u})\big|_{\boldsymbol{u}=\boldsymbol{t}_{\boldsymbol{y}}}$ of the estimator $\hat{\theta}_j := h(\hat{\boldsymbol{t}}_{\boldsymbol{y},j})$ in each iteration of the active sampling algorithm. We now show how this theoretically could be achieved under the multinomial sampling design. A general result is provided in Proposition 1. Corresponding optimal designs for estimating a total t_y or mean t_y/t_r are presented in Corollary 1.

Proposition 1 *Let* $\boldsymbol{S}_j = (S_{j1}, \ldots, S_{jN})$, $j \leq k$, $\boldsymbol{S}_k|\boldsymbol{S}_1, \ldots, \boldsymbol{S}_{k-1} \sim \mathrm{Multinomial}(n_k, \boldsymbol{\pi}_k)$ *and* $w_{ki} := \mathrm{E}[S_{ki}|\boldsymbol{S}_1, \ldots, \boldsymbol{S}_{k-1}]^{-1} = (n_k \pi_{ki})^{-1}$. *Consider a characteristic* $\theta := h(\boldsymbol{t}_{\boldsymbol{y}})$ *for some function* $h : \mathbb{R}^d \to \mathbb{R}$ *and* d*-dimensional vector of totals* $\boldsymbol{t}_{\boldsymbol{y}}$*, and assume that the function* h *is differentiable in a neighborhood of* $\boldsymbol{t}_{\boldsymbol{y}}$ *with* $\nabla h(\boldsymbol{u})\big|_{\boldsymbol{u}=\boldsymbol{t}_{\boldsymbol{y}}} \neq \mathbf{0}$. *Let*

$$\hat{\boldsymbol{t}}_{\boldsymbol{y},k} = \sum_{i=1}^{N} S_{ki} w_{ki} p_i r_i \boldsymbol{y}_i, \quad \textit{and} \quad \boldsymbol{\Phi}_k = \mathbf{Cov}(\hat{\boldsymbol{t}}_{\boldsymbol{y},k}|\boldsymbol{S}_1, \ldots, \boldsymbol{S}_{k-1}).$$

As a function of the sampling scheme $\boldsymbol{\pi}_k = (\pi_{k1}, \ldots, \pi_{kN})$*, the approximate variance* $\mathrm{AV}(\hat{\theta}_k) := \nabla h(\boldsymbol{u})^T \boldsymbol{\Phi}_k \nabla h(\boldsymbol{u})\big|_{\boldsymbol{u}=\boldsymbol{t}_{\boldsymbol{y}}}$ *of the estimator* $\hat{\theta}_k := h(\hat{\boldsymbol{t}}_{\boldsymbol{y},k})$ *is minimized by*

$$\boldsymbol{\pi}_k^* = (\pi_{k1}^*, \ldots, \pi_{kN}^*), \quad \pi_{ki}^* = \frac{p_i r_i v_i}{\sum_{j=1}^N p_j r_j v_j}, \tag{5}$$

with $v_i = \left|\nabla h(\boldsymbol{u})^T \boldsymbol{y}_i\right|_{\boldsymbol{u}=\boldsymbol{t_y}}$.

Corollary 1 *Let* $\boldsymbol{S}_k = (S_{k1}, \ldots, S_{kN})$ *and* w_{ki} *be defined according to Proposition 1. Let*

$$\hat{t}_{y,k} = \sum_{i=1}^{N} S_{ki} w_{ki} p_i r_i y_i, \quad \textit{and} \quad \hat{t}_{r,k} = \sum_{i=1}^{N} S_{ki} w_{ki} p_i r_i.$$

Then, as a function of the sampling scheme $\boldsymbol{\pi}_k = (\pi_{k1}, \ldots, \pi_{kN})$*:*

a) The variance of $\hat{t}_{y,k}$ *is minimized by* (5) *with* $v_i = |y_i|$.

b) The approximate variance of $\hat{t}_{y,k}/\hat{t}_{r,k}$ *is minimized by* (5) *with* $v_i = |y_i - t_y/t_r|$*, provided that* $t_r > 0$.

Corollary 1 a) is a standard result: for linear statistics, instances should be sampled with probabilities proportional to 'size', in this case $p_i r_i |y_i|$. Generally, we conclude that instances should be assigned probabilities proportional to 'influence', as given by the gradient $p_i r_i \left|\nabla h(\boldsymbol{u})^T \boldsymbol{y}_i\right|_{\boldsymbol{u}=\boldsymbol{t_y}}$. To see this, note that $\left|\nabla h(\boldsymbol{u})^T \boldsymbol{y}_i\right|_{\boldsymbol{u}=\boldsymbol{t_y}}$ is large when $\boldsymbol{y}_i$ is large (in Euclidean norm), and aligned with the gradient $\nabla h(\boldsymbol{u})$ (i.e., the direction of steepest change) at the true population parameter $\boldsymbol{t_y}$. Such instances will have a large influence on estimation, and should according to Proposition 1 be oversampled for optimal performance.

4.2.2 Optimal auxiliary-information-assisted sampling schemes

Unfortunately, the results of Proposition 1 and Corollary 1 are not of immediate practical use, since evaluating the theoretically optimal sampling schemes requires knowledge about the yet unobserved variables r_i and $\boldsymbol{y}_i$. Hence, we treat the unknown values of the relevance indicator variables r_i and outcomes $\boldsymbol{y}_i$ as random variables $(R_i, \boldsymbol{Y}_i^T)$, and include in our active sampling algorithm an auxiliary model $f(r_i, \boldsymbol{y}_i^T | \boldsymbol{z}_i^T)$ for the joint distribution of $(R_i, \boldsymbol{Y}_i^T)$ given auxiliary variables $\boldsymbol{z}_i$. This model is intended to capture our predictions and uncertainties in yet unseen data. In the spirit of traditional importance sampling methods and probability-proportional-to-size sampling, a first naive attempt towards an optimal active sampling method would be to plug in the current estimate $\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k-1)}$ and predictions $(\hat{r}_i, \hat{\boldsymbol{y}}_i^T)$ of $(r_i, \boldsymbol{y}_i^T)$ into the formulas of Proposition 1/Corollary 1 to calculate the sampling scheme for the next iteration of the algorithm (Subroutine 1, Method 1). However, this treats the predictions as the true values and ignores the uncertainty of the predictions. Indeed, there are generally many different values of $(r_i, \boldsymbol{y}_i^T)$ compatible with existing auxiliary information $\boldsymbol{z}_i$. To derive sampling schemes with good performance, we need to account for this uncertainty.

Following Isaki and Fuller (1982), we define the anticipated variance of a statistic $\hat{\theta}$ as

$$\mathrm{E}_{\boldsymbol{R},\boldsymbol{Y}}[\mathrm{Var}(\hat{\theta}|\boldsymbol{R}=\tilde{\boldsymbol{r}},\boldsymbol{Y}=\tilde{\boldsymbol{y}})],$$

where the inner term denotes the (approximate) variance of $\hat{\theta}$ with respect to the sampling mechanism given data $\{(\tilde{r}_i,\tilde{\boldsymbol{y}}_i^T)\}_{i\in\mathcal{D}}$ generated according to $f(r_i,\boldsymbol{y}_i^T|\boldsymbol{z}_i^T)$, and the outer term denotes expectation with respect to the random variables $\{(R_i,\boldsymbol{Y}_i^T)\}_{i\in\mathcal{D}}$. We may think of this as the expected variance due to subsampling from a dataset $\tilde{\mathcal{D}}$, averaged over all possible datasets $\tilde{\mathcal{D}}$ compatible with the auxiliary information $\{\boldsymbol{z}_i\}_{i\in\mathcal{D}}$ under the model $f(r_i,\boldsymbol{y}_i^T|\boldsymbol{z}_i^T)$. In other words, the anticipated variance is our prediction of the actual (unknown) variance of the statistic $\hat{\theta}$. We derive sampling schemes with improved performance by minimizing the anticipated variance of our estimator under an assisting auxiliary model $f(r_i,\boldsymbol{y}_i^T|\boldsymbol{z}_i^T)$. A general result is provided in Proposition 2. Corresponding optimal designs for estimating a total t_y or mean t_y/t_r are presented in Corollary 2.

Proposition 2 *Let $\boldsymbol{S}_k$, $\hat{\boldsymbol{t}}_{\boldsymbol{y},k}$, $\boldsymbol{\Phi}_k$, and function $h:\mathbb{R}^d\to\mathbb{R}$ be defined according to Proposition 1. Assume that $\{R_i\}_{i=1}^N$ is a collection of independent* Bernoulli($\hat{r}_i$) *random variables, and $\{\boldsymbol{Y}\}_{i=1}^N$ a collection of independent random vectors with $\mathrm{E}[\boldsymbol{Y}_i|R_i=1]=\hat{\boldsymbol{y}}_i$ and $\mathbf{Cov}(\boldsymbol{Y}_i|R_i=1)=\boldsymbol{\Sigma}_i$. Let $\boldsymbol{R}=(R_1,\ldots,R_N)^T$, $\boldsymbol{Y}=(\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_N)^T$. As a function of the sampling scheme $\boldsymbol{\pi}_k=(\pi_{k1},\ldots,\pi_{kN})$, the anticipated approximate variance $\mathrm{E}_{(\boldsymbol{R},\boldsymbol{Y})}[\nabla h(\boldsymbol{u})^T\boldsymbol{\Phi}_k\nabla h(\boldsymbol{u})\big|_{\boldsymbol{u}=\boldsymbol{t}_{\boldsymbol{y}}}$ of the estimator $\hat{\theta}_k:=h(\hat{\boldsymbol{t}}_{\boldsymbol{y},k})$ is minimized by*

$$\boldsymbol{\pi}_k^*=(\pi_{k1}^*,\ldots,\pi_{kN}^*),\quad \pi_{ki}^*=\frac{p_i\sqrt{v_i}}{\sum_{j=1}^N p_j\sqrt{v_j}},\tag{6}$$

with

$$v_i=\hat{r}_i\left[(\nabla h(\boldsymbol{u})^T\hat{\boldsymbol{y}}_i)^2+\nabla h(\boldsymbol{u})^T\boldsymbol{\Sigma}_i\nabla h(\boldsymbol{u})\right]\big|_{\boldsymbol{u}=\boldsymbol{t}_{\boldsymbol{y}}}.$$

Corollary 2 *Let $\boldsymbol{S}_k$, $\hat{t}_{y,k}$, and $\hat{t}_{r,k}$ be defined according to Corollary 1. Assume that $\{R_i\}_{i=1}^N$ is a collection of independent* Bernoulli($\hat{r}_i$) *random variables, and $\{Y_i\}_{i=1}^N$ a collection of independent random variables with $\mathrm{E}(Y_i|R_i=1)=\hat{y}_i$ and $\mathrm{Var}(Y_i|R_i=1)=\sigma_i^2$. Then, as a function of the sampling scheme $\boldsymbol{\pi}_k=(\pi_{k1},\ldots,\pi_{kN})$:*

a) The anticipated variance of $\hat{t}_{y,k}$ is minimized by (6) *with $v_i=\hat{r}_i(\hat{y}_i^2+\sigma_i^2)$.*

b) The anticipated approximate variance of $\hat{t}_y/\hat{t}_r$ is minimized by (6) *with $v_i=\hat{r}_i\left[(\hat{y}_i-t_y/t_r)^2+\sigma_i^2\right]$, provided that $t_r>0$.*

We note that the optimal sampling schemes of Proposition 2 and Corollary 2 only depend on the first two moments of the unknowns $\boldsymbol{Y}_i$. Hence, our results are immediately applicable

to any type of random variables, including binary and discrete as well as continuous variables. We also note that the result of Proposition 2/Corollary 2 coincides with that of Proposition 1/Corollary 1 when $(r_i, \boldsymbol{y}_i^T)$ are predictable without error, i.e., when all $\hat{r}_i = r_i$ and $\sigma_i = 0$. Compared to the naive approach of plugging in predictions to the sampling schemes of Proposition 1/Corollary 1, the sampling schemes in Proposition 2/Corollary 2 explicitly account for prediction uncertainty through the inclusion of residual variances σ_i^2. In the homoscedastic case, i.e., when all σ_i are equal, this results in a regularization of the sampling scheme of the plug-in approach towards a density sampling scheme with probabilities proportional to the prior observation weights p_i, or towards simple random sampling if all prior weights p_i are equal. Moreover, the amount of regularization is determined by the prediction uncertainty and hence is completely data-driven.

When implementing the sampling schemes of Proposition 2/Corollary 2 in the active sampling algorithm, we replace the population quantities, predictions and model parameters by their corresponding estimates and predictions based on currently available data (Subroutine 1, Method 2). As functions of random variables, the sampling schemes are also subject to random variation. This may cause unstable performance due to incidentally large sampling weights, particularly in early iterations and small samples. Care should therefore be taken to avoid overfitting in the learning step preceding the optimization of the sampling scheme in the active sampling algorithm. It is also important that an unbiased estimate of the residual variance is used. Such an estimate may be obtained by evaluation of the predictions on hold-out data, using, e.g., cross-validation. Underestimation of the residual variance and overoptimism in the predictions may result in sampling probabilities too close to zero, with highly variable sampling weights and sub-optimal performance as a result.

5 Application and empirical evaluation

We evaluated the empirical performance of the active sampling method on the crash-causation-based scenario generation problem introduced in Section 2. The empirical evaluation was conducted by repeated subsampling from a large dataset, denoted as the ground truth dataset, pretending that only small subset of the instances in this dataset could be fully observed. Section 5.1 introduces the data, model and simulation setup, together with methods for performance evaluation. Empirical results are presented in Section 5.2.

5.1 Data and Methods

Scenario generation framework. The data model is based on the generation of crash scenarios through virtual simulations of a set of reconstructed rear-end crashes, to which a crash-causation model and a driver response model is applied. The crash-causation model consists of two components: one based on drivers not keeping their eyes on the forward roadway, and the other based on the fact that drivers do not brake at the maximum level (i.e., the performance limit of the vehicle and roadway) even if they are about to crash. The simulations for this work were run with the kinematics of the lead vehicle of each original (real) rear-end crash and using crash-causation and response models to replace the evasive maneuver of the following vehicle. This was done by simulating the (counterfactual) outcome if the glances off road and deceleration of the following vehicle had been different from each original crash. Crashes occur in the simulations under certain combinations of off-road glances and driver maximum deceleration. The longer the off-road glance duration and the lower the deceleration, the higher the probability of a crash and the higher the impact speed if there is a crash.

Ground truth dataset. The data used for scenario generation in this study were reconstructed pre-crash kinematics of 44 rear-end crashes from a crash database provided by Volvo Car Corporation. This database contains information about crashes that occurred with Volvo vehicles in Sweden (Coelingh et al., 2007). We constructed a ground truth dataset by running virtual simulations for all 1005 combinations of glance duration (67 levels, 0.0–6.6s) and deceleration (15 levels, 3.3–10.3 m/s^2) for all 44 crashes. The simulations were run under both manual driving (baseline scenario) and automated emergency braking (AEB) system conditions, producing a dataset of 44220 pairs of observations.

The outputs of the simulations were the impact speed under both scenarios (baseline and AEB). We also calculated the impact speed reduction (continuous) and crash avoidance (binary) of the AEB system compared to the baseline scenario. The aim in our experiments was to estimate the benefit of the AEB system, as measured by mean impact speed reduction and crash avoidance rate compared to manual baseline driving, given that there was a crash in the baseline scenario.

Performance evaluation. We evaluated the properties and performance of the active sampling method by repeated subsampling from the ground truth dataset. The following properties, performance measures and comparisons were considered:

i) **Asymptotic normality and confidence interval coverage**: We evaluated the

coverage rates of large-sample normal confidence intervals (4) with the three different methods for variance estimation described in Section 4.1: the classical method (Sen-Yates-Grundy estimator), martingale method, and the bootstrap method. This was done for batch sizes of 10, up to a total sample size of 300 observations.

ii) **Comparison of active sampling methods**: We evaluated the performance of active sampling with the two different implementations of active sampling schemes described in Subroutine 1: the naive approach (Method 1) where predictions are inserted immediately into the theoretically optimal designs of Corollary 1 b), and the anticipated-variance-minimizing scheme (Method 2) which additionally accounts for prediction uncertainty according to Corollary 2 b).

iii) **Active sampling compared to traditional methods**: We evaluated the performance of active sampling optimized for estimating the mean impact speed reduction or crash avoidance rate of an AEB system compared to baseline driving (without AEB), vs. simple random sampling and importance sampling. Two importance sampling schemes were considered: a density sampling scheme with probabilities proportional to the prior observation weights p_i, and a severity sampling scheme that additionally attempts to oversample high-severity instances.

For ii) and iii), performance was measured as the root mean squared error (RMSE) from ground truth when estimating the mean impact speed reduction and crash avoidance rate of the AEB system compared to baseline driving. Each sampling method was repeated 300 times for sample sizes up to $n = 2000$ observations, and the average performance evaluated. The results are presented graphically as functions of the sample size, i.e., the number of baseline-AEB simulations pairs. Further details are provided in Appendix B.

Implementation. Active sampling was implemented according to Algorithm 1, with batch sizes of $n_k = 10$ observations per iteration and sampling schemes calculated according to the anticipated-variance-minimizing scheme in Corollary 2 b) (Subroutine 1, Method 2), unless otherwise stated. The empirical evaluation was implemented using the `R` language and environment for statistical computing, version 4.2.1 (R Core Team, 2022). For the learning step of the active sampling algorithm, we used the random forest method (Breiman, 2001) as implemented in the `ranger` package version 0.14.1 (Wright and Ziegler, 2017), with hyper-parameter tuning by cross-validation using the `caret` package version 6.0-92 (Kuhn, 2022). Bootstrap variance estimation was performed with 500 replicates and implemented using the `R boot` package version 1.3-28 (Canty and Ripley, 2021). The complete `R` code for

the active sampling algorithm and simulation experiments and data are available online at `https://github.com/imbhe/ActiveSampling`. Further implementation details are also provided in Appendix B.

5.2 Results of empirical evaluation

Confidence interval coverage rates. The empirical coverage rates of large sample normal confidence intervals under active sampling with three different methods for variance estimation are presented in Figure 1. There was a clear under-coverage in small samples, as expected. Both the classical variance estimator and bootstrap method produced confidence intervals that approached the nominal 95% confidence level relatively quickly as the sample size increased. Coverage rates were somewhat lower with the martingale method, and more iterations where needed before the nominal 95% level was reached. Estimating the crash avoidance rate (mean of a binary variable) required more samples than estimating the mean impact speed reduction (mean of a continuous variable) to reach the nominal 95% coverage level.

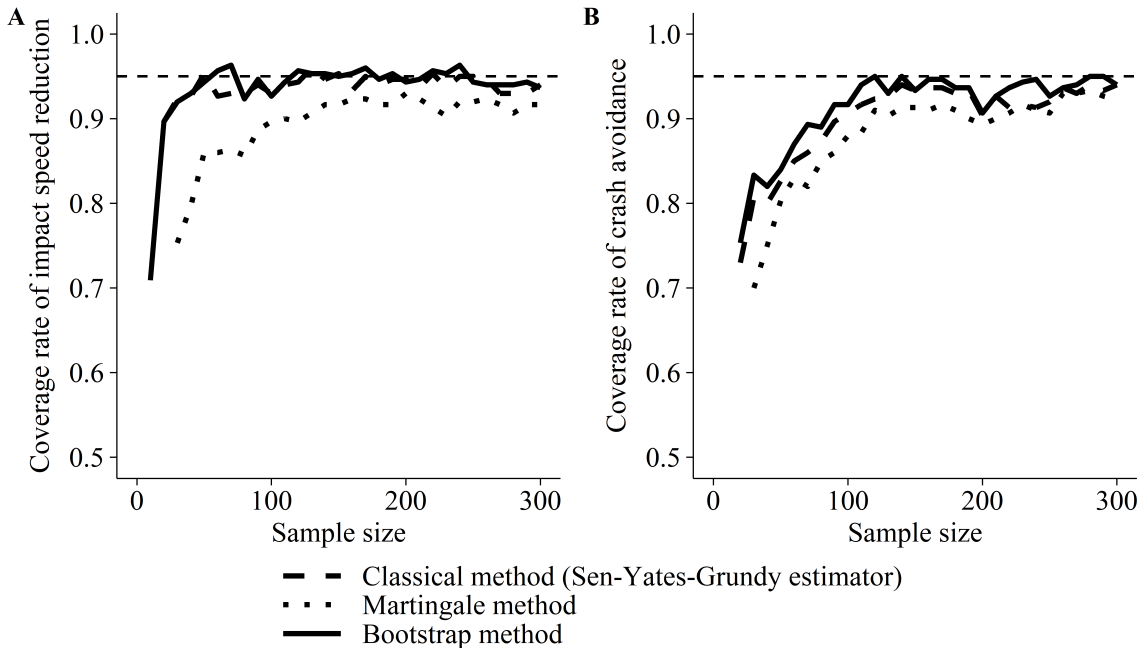

Figure 1: Empirical coverage rates of 95% confidence intervals vs. sample size for (A) mean impact speed reduction and (B) crash avoidance rate. The lines show the coverage rates with three different methods for variance estimation in 300 repeated active sampling experiments.

Comparison of active sampling schemes. Figure 2 shows the RMSEs in estimating the mean impact speed reduction and crash avoidance rate with two different implementations of active sampling schemes. The naive approach, where predictions are inserted immediately into the formulas for the theoretically optimal design in Corollary 1 b), had a poor performance. Substantial improvements were observed with the anticipated-variance-minimizing sampling schemes of Corollary 2 b), i.e., when accounting for prediction uncertainty in the optimization of the sampling schemes.

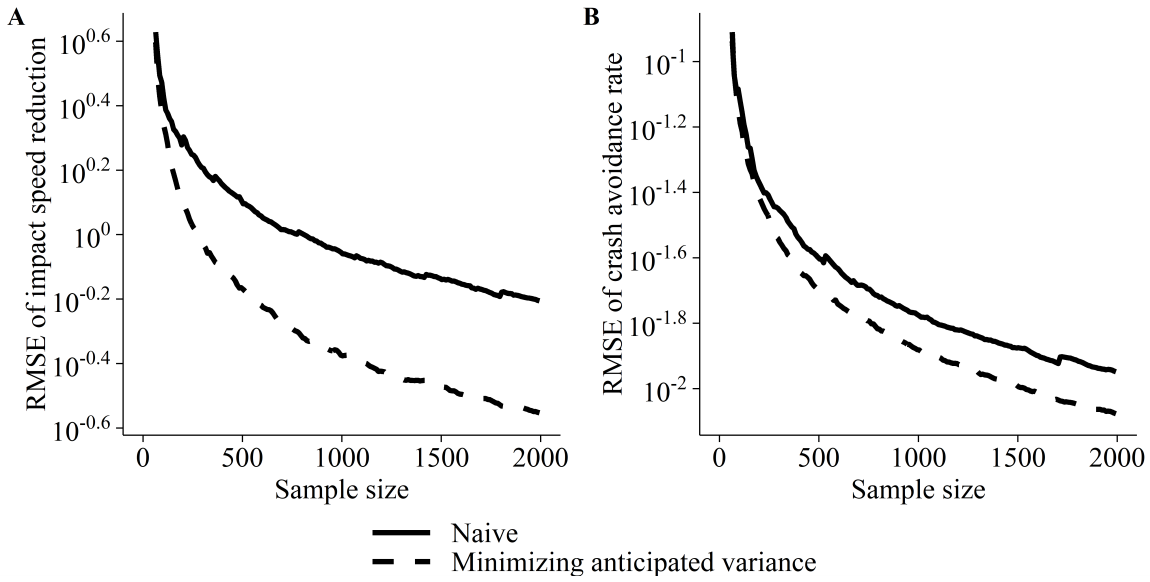

Figure 2: Root mean squared error (RMSE) vs. sample size with active sampling optimized on (A) mean impact speed reduction and (B) crash avoidance rate. The lines show the performance of two different implementations of active sampling schemes: the naive method (Subroutine 1, Method 1) and the anticipated-variance-minimizing method (Subroutine 1, Method 2).

Active sampling vs. traditional sampling methods. The RMSE in estimation with active sampling compared to simple random sampling and traditional importance sampling methods is presented in Figure 3. As expected, simple random sampling had the worst performance. The two importance sampling schemes had similar performance, with a slight advantage of severity importance sampling for estimating the crash avoidance rate. Active sampling optimized for a specific characteristic always had best performance on the characteristic for which it was optimized. For sample sizes $n \geq 500$, active sampling required 7.2–47.6% less observations than importance sampling to reach the same level of performance on the characteristic for which it was optimized. The benefit of active sampling increased with the sample size. At $n = 2000$ observations, we observed a reduction in

RMSE of 21.7–37.9% with active sampling compared to importance sampling. Moreover, active sampling performance was on par with that of traditional methods when evaluated on characteristics other than the one it was optimized for.

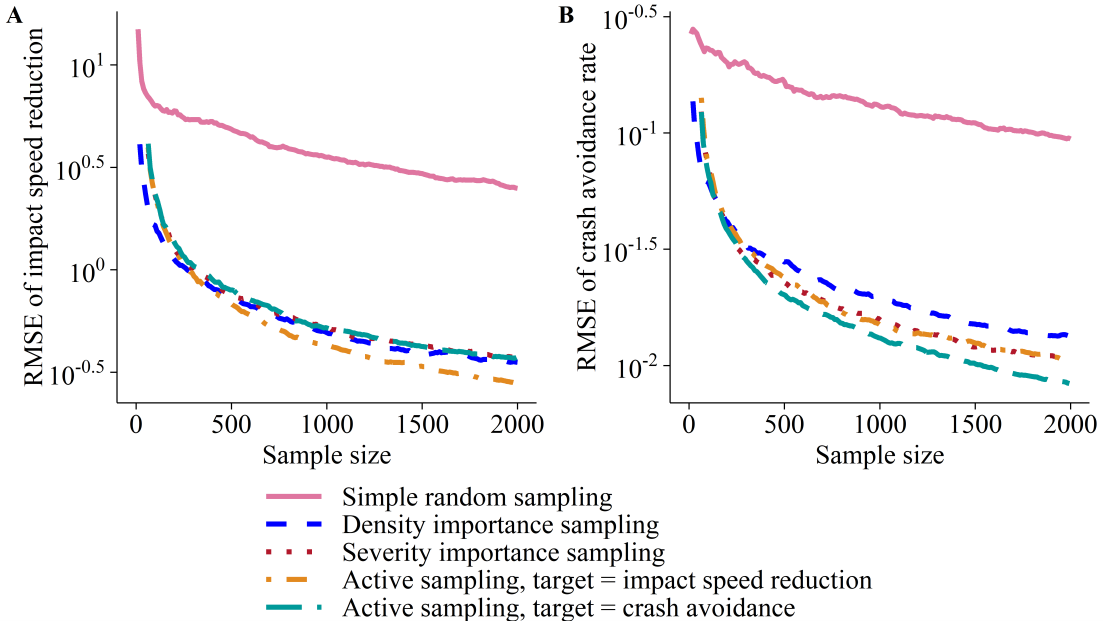

Figure 3: Root mean squared error (RMSE) vs. sample size in the estimation of (A) the mean impact speed reduction and (B) crash avoidance rate. The lines show the performance using simple random sampling, importance sampling, and active sampling optimized for the estimation of mean impact speed reduction and crash avoidance rate.

6 Discussion

We have introduced an active sampling framework for optimal sampling and estimation of finite population characteristics in measurement-constrained experiments. Inspired by active learning, the method iterates between parameter estimation and data collection by adaptive importance sampling with optimal subsamples guided by machine learning predictions on yet unseen data. Active sampling overcomes the limitations of traditional importance sampling methods in terms of prior knowledge requirements and manual input to the construction of sampling schemes, and offers a highly flexible and completely data-driven procedure to sample selection. We have evaluated the performance of active sampling for safety assessment of advanced driver assistance systems in the context of crash-causation-based scenario generation. Substantial improvements over traditional importance sampling

methods were demonstrated, with sample size reductions of up to 50% for the same level of performance in terms of RMSE.

We have conducted an asymptotic analysis of the properties of the active sampling method, and proved theoretically that active sampling under mild assumptions produces consistent and asymptotically normally distributed estimators. Our theoretical results were also confirmed empirically in our experiments. The two major assumptions were i) bounded second moments on the random variables involved, and ii) asymptotically negligible dependencies between the iterations of the active sampling algorithm. The first of these may be justified by ensuring that the sampling probabilities are properly bounded away from zero. In our algorithm this is automatically achieved by accounting for residual uncertainty when calculating the optimal sampling scheme. The second assumption may be justified by designing active sampling strategies so that each individual observation has a limited influence on the sampling schemes and selections in future iterations. This is accomplished in our algorithm by calculating sampling schemes based on certain functions of empirical characteristics of the data. See also Bach (2007) for a related discussion in the active learning context. Although we were primarily concerned with the case where the batch sizes n_k are small and the number of iterations k large, we conjecture that similar results also hold true in the case where all $n_1, \ldots, n_k$ tend to infinity (at same rate) with k fixed.

Three different methods for variance estimation were proposed: a classical method, which uses a pooled estimator of the conditional variances in each iteration of the algorithm, and the conditional variances are estimated using classical survey sampling techniques; a martingale method, which uses the squared variation of the estimates in the different iterations of the active sampling algorithm to estimate the total variance; and a simple non-parametric bootstrap method. Both the classical and bootstrap method performed well already at small samples. Indeed, these are accurate of the order $O(m_k^{-1})$, where m_k is the total sample size after k iterations. In contrast, the martingale method is accurate of the order $O(k^{-1})$, and hence requires more iterations to provide reliable estimates. The martingale method should therefore primarily be used when the batch sizes n_k are small and number of iterations k large. Notably, the classical method is not applicable for fixed-size designs when $n_k = 1$, in which case the martingale method is preferred. Both the classical and martingale method may be used internally in the algorithm to monitor the precision and determine when to stop, whereas the bootstrap method, due to its increased computational complexity, is better suited for use after subsampling has been completed. In our experience, all methods produce similar estimates when the assumptions are fulfilled.

Our empirical experiments evaluated the performance of two different implementations of

active sampling, two importance sampling methods, and simple random sampling. Among these, simple random sampling had the worst performance. This can be explained by it's ignorance to the prior observation weights p_i. Indeed, density importance sampling (probability-proportional-to-size sampling) is a better choice when such observation weights are present. The naive implementation of active sampling, where predicted values are inserted immediately into the formulas of the theoretically optimal design, also had a poor performance. Nonetheless, similar approaches are often suggested in various subsampling applications (c.f Chu et al., 2011; Ganti and Gray, 2012; Farquhar et al., 2021). Our theoretical analysis suggests that this is sub-optimal, as it fails to account for prediction uncertainty. In contrast, the anticipated-variance-minimizing active sampling scheme, which accounts for prediction uncertainty, had substantial improvements over traditional importance sampling.

A well-known issue with methods based on inverse probability weighting is the risk of variance inflation due to incidentally large sampling weights. Hence, a common suggestion in the subsampling literature is to put a lower limit on the sampling probabilities and regularize the sampling scheme towards a more uniform scheme (c.f Chu et al., 2011; Ganti and Gray, 2012; Ma et al., 2015). However, such adjustments are often introduced ad hoc, with additional hyper-parameters to be specified. We note that our anticipated-variance-minimizing active sampling scheme achieves a similar effect in a completely data-driven manner. Hence, our results provide theoretical arguments to why such regularization may be beneficial, reveal how it is related to prediction uncertainty, and show how it should be implemented to achieve optimal performance. Empirical results were consistent with these assertions.

This paper illustrated the active sampling method in an application to generation of simulation scenarios for the assessment of automated emergency braking. Not only can the method be applied more broadly in the traffic safety domain, such as for virtual safety assessment of self-driving vehicles of the future, but it can be applied to a wide range of subsampling applications. Future research on the topic may pursue more efficient methods of partitioning the dataset into areas where the outcomes are more precisely predicted or known (where subsampling is less useful) and those where outcomes are less precisely predicted, as well as demonstrate practical applications further.

7 Conclusion

We have introduced a machine-learning-assisted active sampling framework for optimal sampling and inference for finite populations and massive datasets. Methods for variance and interval estimation have been proposed, and their validity in the active sampling setting was confirmed empirically. Properly accounting for prediction uncertainty was crucial for the performance of the active sampling algorithm. Substantial performance improvements were observed compared to traditional importance sampling methods. Active sampling is a promising method for efficient sampling and inference in subsampling applications.

Acknowledgment

We would like to thank Volvo Car Corporation for allowing us to use their data and simulation tool, and in particular Malin Svärd and Simon Lundell at Volvo for supporting in the simulation setup. We further want to thank Marina Axelson-Fisk and Johan Jonasson for valuable comments on the manuscript.

Funding

This research was supported by the European Commission through the SHAPE-IT project under the European Union's Horizon 2020 research and innovation programme (under the Marie Skłodowska-Curie grant agreement 860410), and in part also by the Swedish funding agency VINNOVA through the FFI project QUADRIS. Also Chalmers Area of Advance Transport funded part of this research.

SUPPLEMENTARY MATERIAL

Additional theoretical material and implementation details: Additional theoretical results and proofs (Appendix A), and additional details on the implementation of the sampling methods in the empirical evaluation (Appendix B).

R Code for empirical evaluation: Code and data used for the empirical evaluation in Section 5 is available online at `https://github.com/imbhe/ActiveSampling`.

References

Akagi, Y., Kato, R., Kitajima, S., Antona-Makoshi, J., and Uchida, N. (2019). A risk-index based sampling method to generate scenarios for the evaluation of automated driving vehicle safety. In *2019 IEEE Intelligent Transportation Systems Conference*.

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*.

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

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

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

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

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

Bärgman, J., Svärd, M., Lundell, S., and Shams El Din, A. (2022). Validation of an eyes-off-road crash causation model for virtual safety assessment. In *Proceedings of the 8th International Conference on Driver Distraction and Inattention*.

Canty, A. and Ripley, B. D. (2021). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-28.

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

Coelingh, E., Jakobsson, L., Lind, H., and Lindman, M. (2007). Collision warning with auto brake: A real-life safety perspective. *Innovations for Safety: Opportunities and Challenges*.

Dai, W., Song, Y., and Wang, D. (2022). A subsampling method for regression problems based on minimum energy criterion. *Technometrics*, 0(0):1–14.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge, UK.

de Gelder, E. and Paardekooper, J.-P. (2017). Assessment of automated driving systems using real-life scenarios. In *2017 IEEE Intelligent Vehicles Symposium*.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26.

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

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

Ganti, R. and Gray, A. (2012). UPAL: Unbiased pool based active learning. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*.

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

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.

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.

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. (2022). Optimization of two-phase sampling designs with application to naturalistic driving studies. *IEEE Transactions on Intelligent Transportation Systems*, 23(4):3575–3588.

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.

Kossen, J., Farquhar, S., Gal, Y., and Rainforth, T. (2021). Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning*.

Kuhn, M. (2022). *caret: Classification and Regression Training*. R package version 6.0-92.

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. and Sun, X. (2015). Leveraging for big data regression. *WIREs Computational Statistics*, 7(1):70–76.

Magnusson, M., Andersen, M., Jonasson, J., and Vehtari, A. (2019). Bayesian leave-one-out cross-validation for large data. In *Proceedings of the 36th International Conference on Machine Learning*.

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.

Mousa, S. R., Bakhit, P. R., and Ishak, S. (2019). An extreme gradient boosting method for identifying the factors contributing to crash/near-crash events: A naturalistic driving study. *Canadian Journal of Civil Engineering*, 46(8):712–721.

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

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

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sander, U. (2018). *Predicting Safety Benefits of Automated Emergency Braking at Intersections-Virtual Simulations Based on Real-World Accident Data*. PhD thesis, Chalmers University of Technology, Gothenburg, Sweden.

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

Sen, P. and Singer, J. (1993). *Large Sample Methods in Statistics: An Introduction with Applications*. CRC Press, Boca Raton, FL.

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*, 2021.

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

Tille, Y. (2020). *Sampling and Estimation from Finite Populations*. Wiley, Hoboken, NJ.

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, X., Peng, H., and Zhao, D. (2021). Combining reachability analysis and importance sampling for accelerated evaluation of highway automated vehicles at pedestrian crossing. *ASME Letters in Dynamic Systems and Control*, 1(1):011017.

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.

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

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.

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

Zhang, X., Tao, J., Tan, K., Törngren, M., Sánchez, J. M. G., Ramli, M. R., Tao, X., Gyllenhammar, M., Wotawa, F., Mohan, N., et al. (2021). Finding critical scenarios for automated driving systems: A systematic literature review. *arXiv:2110.08664*.

A Additional theoretical results and proofs

This appendix contains additional theoretical results and proofs. An asymptotic analysis of the active sampling estimators is presented in Appendix A.1. Proofs of the optimality results of Proposition 1–2 and Corollary 1–2 are presented in Appendix A.2.

A.1 Central limit theorems

We provide in Proposition S1 conditions under which the active sampling estimator $\hat{t}_y^{(k)}$ of a total t_y is consistent and asymptotically normally distributed, and present consistent variance estimators. A generalization to multivariate estimators and to characteristics defined as smooth functions of totals is provided in Corollary S1. For a sequence of random variables $\{X, X_n, n \geq 1\}$, we will use $X_n \xrightarrow{d} X$, $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{L_r} X$ to denote convergence of X_n to X in distribution, probability, and r^{th} mean, respectively. To show asymptotic normality of our active sampling estimator, we use the following result of Brown (1971):

Lemma 1 (Martingale central limit theorem)
Consider a sequence $\{X_j\}_{j=1}^{\infty}$ *of random variables such that* $\mathrm{E}[X_j] = \mathrm{E}[X_j|X_1, \ldots, X_{j-1}] = 0$ *and* $\mathrm{E}[X_j^2] < \infty$. *Let* $\sigma_j^2 = \mathrm{E}[X_j^2|X_1, \ldots, X_{j-1}]$, $U_k = \sum_{j=1}^k X_j$, $V_k^2 = \sum_{j=1}^k \sigma_j^2$, *and* $u_k^2 = \mathrm{E}[U_k^2] = \mathrm{E}[V_k^2]$. *Assume that* $V_k^2 u_k^{-2} \xrightarrow{p} 1$ *as* $k \to \infty$, *and that the Lindeberg-Feller condition holds:*

$$u_k^{-2} \sum_{j=1}^{k} \mathrm{E}[X_j^2 I(|X_j| > \varepsilon u_k)] \to 0 \quad \text{as } k \to \infty \quad \text{for all } \varepsilon > 0. \tag{S.1}$$

Then

$$U_k/u_k \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as } k \to \infty.$$

We will also make use of the following results:

i) **Dominated convergence theorem**: Let $\{X, X_n, n \geq 1\}$ be a sequence of random variables such that $X_n \xrightarrow{p} X$ and $\mathrm{E}[\sup_{j\geq 1} |X_j|] < \infty$. Then $X_n \xrightarrow{L_1} X$.

ii) **Cramér-Wold theorem**: Let $\boldsymbol{X}, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be random vectors in $\mathbb{R}^d$. Then $\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{X}$ if and only if, for every fixed $\boldsymbol{\lambda} \in \mathbb{R}^d$, we have $\boldsymbol{\lambda}^T \boldsymbol{X}_n \xrightarrow{d} \boldsymbol{\lambda}^T \boldsymbol{X}$.

iii) **Delta method**: Let $\{\boldsymbol{X}_n\}$ be a sequence of random vectors such that $\sqrt{n}(\boldsymbol{X}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Sigma})$. Consider a function $h : \mathbb{R}^d \to \mathbb{R}$ and assume that $h(\boldsymbol{u})$ is differentiable in a neighborhood of $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, with $\nabla h(\boldsymbol{u})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \neq \boldsymbol{0}$. Then

$$\sqrt{n}(h(\boldsymbol{T}_n) - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \gamma^2), \quad \text{with} \quad \gamma^2 = \nabla h(\boldsymbol{\theta})^T \boldsymbol{\Sigma} \nabla h(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

Details may be found in any textbook on large sample methods in statistics (see, e.g., Sen and Singer, 1993). A central limit theorem for active sampling is presented below.

Proposition S1 (Central limit theorem, n_k bounded, $k \to \infty$)
Consider a finite index set $\mathcal{D} = \{1, \ldots, N\}$ with corresponding data $y_1, \ldots, y_N$, and infinite sequence $\{n_j\}_{j=1}^{\infty}$ with $n_j \in \mathbb{N}$, $n_j < N$. Let $\{\boldsymbol{S}_j\}_{j=1}^{\infty}$ be an infinite sequence of random vectors $\boldsymbol{S}_j = (S_{j1}, \ldots, S_{jN}) \in \mathbb{N}^N$ such that $\sum_{i=1}^{N} \mu_{ji} = n_j$, where $\mu_{ji} := \mathrm{E}[S_{ji} | \boldsymbol{S}_1, \ldots \boldsymbol{S}_{j-1}]$ are assumed to be strictly positive for all j, i. Let $t_y = \sum_{i \in \mathcal{D}} y_i$, $\hat{t}_{y,j} = \sum_{i \in \mathcal{D}} \frac{S_{ji} y_i}{\mu_{ji}}$, $m_k = \sum_{j=1}^{k} n_j$, $\hat{t}_y^{(k)} = \frac{1}{m_k} \sum_{j=1}^{k} n_j \hat{t}_{y,j}$, $\sigma_j^2 = \mathrm{Var}(\hat{t}_{y,j} | \boldsymbol{S}_1, \ldots, \boldsymbol{S}_{j-1})$, $A_k^2 = \sum_{j=1}^{k} n_j^2 \sigma_j^2$, and $b_k^2 = \mathrm{Var}(\sum_{j=1}^{k} n_j \hat{t}_{y,j})$. Assume that

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

(A2) $b_k \to \infty$ *as* $k \to \infty$, *and*

(A3) $A_k^2 b_k^{-2} \xrightarrow{p} 1$ *as* $k \to \infty$.

Then

$$\frac{\hat{t}_y^{(k)} - t_y}{b_k / m_k} \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as } k \to \infty, \quad \text{and} \tag{S.2}$$

$$b_k^{-2} \sum_{j=1}^{k} n_j^2 \left(\hat{t}_{y,j} - \hat{t}_y^{(k)} \right)^2 \xrightarrow{p} 1 \quad \text{as } k \to \infty. \tag{S.3}$$

Furthermore, if $\hat{\sigma}_j^2$ are unbiased estimators of the conditional variances σ_j^2, i.e., $\mathrm{E}[\hat{\sigma}_j^2 | \boldsymbol{S}_1, \ldots, \boldsymbol{S}_{j-1}] = \sigma_j^2$, and

(A4) $b_k^{-2} \mathrm{Var}(\sum_{j=1}^{k} n_j \hat{\sigma}_j^2)$ *are uniformly bounded,*

then additionally we have that

$$b_k^{-2} \sum_{j=1}^{k} n_j \hat{\sigma}_j^2 \xrightarrow{p} 1 \quad \text{as } k \to \infty. \tag{S.4}$$

The first result (S.2) establishes asymptotic normality of the active learning estimator $\hat{t}_y^{(k)}$ under the specified conditions. We note that $b_k = O(m_k^{1/2})$, so the convergence of $\hat{t}_y^{(k)}$ to t_y is at the usual parametric rate $m_k^{-1/2}$. The second result (S.3) proves the consistency of the martingale variance estimator $\sum_{j=1}^{k} n_j^2 \left(\hat{t}_{y,j} - \hat{t}_y^{(k)} \right)^2$, and the third (S.4) consistency of the pooled variance estimator $\sum_{j=1}^{k} n_j \hat{\sigma}_j^2$.

Since, in any sensible probability sampling design, S_{ji} have finite second moments, the first assumption (A1) is fulfilled if the sampling probabilities (and corresponding means

μ_{ji}) are properly bounded away from zero. The second assumption (A2) requires the total variance $\mathrm{Var}(\sum_{j=1}^{k} n_j \hat{t}_{y,j})$ to tend to infinity with k. This may at first sight seem to contradict the purpose of active sampling, which is to make the variance as small as possible. Indeed, it is theoretically possible to construct a sampling strategy that produces an estimator with zero variance, which clearly does not converge to a normal limit. In practice, however, finding the true optimal design is not possible and a sampling strategy with good performance generally also fulfills the assumptions (A1) and (A2).

The third assumption states that the sum of conditional variances asymptotically should behave like the total variance. Hence, the statistical properties of the active sampling estimator can be deduced from a single execution of the algorithm. Empirical justification for this assumption is provided in Section 5. We note that the fourth assumption (A4), needed for consistency of the classical variance estimator, is stronger than the second (A2). To see this, note that (A4) requires $\hat{\sigma}_j^2$ to have bounded second moments for every j. But $\hat{\sigma}_j^2$ depends on S_{ji}/μ_{ji}^2, which is larger than S_{ji}/μ_{ji} for all j, i such that $\mu_{ji} \leq 1$, as is the case for all or nearly all j, i in all realistic subsampling applications. For fixed-size designs $\hat{\sigma}_j^2$ also depend on the joint selection probabilities, which means that $\mathrm{E}[S_{ji}S_{jl}]$ need to be properly bounded away from zero for consistent variance estimation. Note, in particular, that this requires all $n_j \geq 2$ for fixed-size designs, whereas (A2) makes no such restriction.

Before providing a proof, we present below a generalization to vectors of totals and smooth functions of totals.

Corollary S1 (Multivariate central limit theorem, n_k bounded, $k \to \infty$) *Consider a finite index set $\mathcal{D} = \{1, \ldots, N\}$ with corresponding data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N \in \mathbb{R}^d$. Let $\{n_j\}_{j=1}^{\infty}$, $\{\boldsymbol{S}_j\}_{j=1}^{\infty}$, μ_{ji} and m_k be defined as in Proposition S1. Let $\boldsymbol{t}_{\boldsymbol{y}} = \sum_{i \in \mathcal{D}} \boldsymbol{y}_i$, $\hat{\boldsymbol{t}}_{\boldsymbol{y},j} = \sum_{i \in \mathcal{D}} \frac{S_{ji}\boldsymbol{y}_i}{\mu_{ji}}$, $\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)} = \frac{1}{m_k} \sum_{j=1}^{k} n_j \hat{\boldsymbol{t}}_{\boldsymbol{y},j}$. Consider a function $h : \mathbb{R}^d \to \mathbb{R}$, and assume that $h(\boldsymbol{u})$ is differentiable in a neighborhood of $\boldsymbol{u} = \boldsymbol{t}_{\boldsymbol{y}}$, with $\nabla h(\boldsymbol{u})|_{\boldsymbol{u}=\boldsymbol{t}_{\boldsymbol{y}}} \neq \boldsymbol{0}$. Let $\boldsymbol{\Phi}_j = \mathbf{Cov}(\hat{\boldsymbol{t}}_{y,j} | \boldsymbol{S}_1, \ldots, \boldsymbol{S}_{j-1})$, $\boldsymbol{A}_k = \sum_{j=1}^{k} n_j^2 \boldsymbol{\Phi}_j$ and $\boldsymbol{B}_k = \mathbf{Cov}(\sum_{j=1}^{k} n_j \hat{\boldsymbol{t}}_{y,j})$. Assume that*

i) S_{ji}/μ_{ji} have uniformly bounded second moments,

ii) $m_k^{-1}\boldsymbol{B}_k$ converges elementwise to some matrix $\boldsymbol{\Psi}$, and $\boldsymbol{\Psi}$ is full rank,

iii) $\boldsymbol{\lambda}^T \boldsymbol{B}_k \boldsymbol{\lambda} \to \infty$ as $k \to \infty$ for every $\boldsymbol{\lambda} \in \mathbb{R}^d \setminus \boldsymbol{0}$, and

iv) $\boldsymbol{A}_k \boldsymbol{B}_k^{-1} \xrightarrow{p} \boldsymbol{I}_{d \times d}$ (elementwise) as $k \to \infty$.

Then

$$\sqrt{m_k}\left(\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)} - \boldsymbol{t}_{\boldsymbol{y}}\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}) \quad \text{as } k \to \infty, \quad \text{and}$$
$$\sqrt{m_k}\left(h(\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)}) - h(\boldsymbol{t}_{\boldsymbol{y}})\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \gamma^2) \quad \text{as } k \to \infty,$$

where $\gamma^2 = \nabla h(\boldsymbol{u})^T \boldsymbol{\Psi} \nabla h(\boldsymbol{u})|_{\boldsymbol{u}=\boldsymbol{t}_{\boldsymbol{y}}}$. *Moreover, the asymptotic covariance matrix* $\boldsymbol{\Psi}$ *and variance* γ^2 *can be consistently estimated by*

$$\hat{\boldsymbol{\Psi}}^{(k)} = \frac{1}{m_k^2} \sum_{j=1}^{k} n_j^2 \left(\hat{\boldsymbol{t}}_{\boldsymbol{y},j} - \hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)}\right)\left(\hat{\boldsymbol{t}}_{\boldsymbol{y},j} - \hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)}\right)^T,$$
$$\hat{\gamma}_k^2 = \nabla h(\boldsymbol{u})^T \hat{\boldsymbol{\Psi}}^{(k)} \nabla h(\boldsymbol{u})\big|_{\boldsymbol{u}=\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)}}.$$

Furthermore, if $\hat{\boldsymbol{\Phi}}_j$ *are unbiased estimators of the conditional covariance matrices* $\boldsymbol{\Phi}_j$, *i.e.,* $\mathrm{E}[\hat{\boldsymbol{\Phi}}_j | \boldsymbol{S}_1, \ldots, \boldsymbol{S}_{j-1}] = \boldsymbol{\Phi}_j$, *and*

iv) $(\boldsymbol{\lambda}^T \boldsymbol{B}_k \boldsymbol{\lambda})^{-1} \mathrm{Var}(\sum_{j=1}^{k} \boldsymbol{\lambda}^T \hat{\boldsymbol{\Phi}}_j \boldsymbol{\lambda})$ *are uniformly bounded for every* $\boldsymbol{\lambda} \in \mathbb{R}^d \setminus \mathbf{0}$,

then the asymptotic covariance matrix $\boldsymbol{\Psi}$ *and variance* γ^2 *can also be consistently estimated by*

$$\widetilde{\boldsymbol{\Psi}}^{(k)} = \frac{1}{m_k^2} \sum_{j=1}^{k} n_j^2 \hat{\boldsymbol{\Phi}}_j,$$
$$\widetilde{\gamma}_k^2 = \nabla h(\boldsymbol{u})^T \widetilde{\boldsymbol{\Psi}}^{(k)} \nabla h(\boldsymbol{u})\big|_{\boldsymbol{u}=\hat{\boldsymbol{t}}_{\boldsymbol{y}}^{(k)}}.$$

Proof *of Proposition S1*
Let $X_j = n_j(\hat{t}_{y,j} - t_y)$, $U_k = \sum_{j=1}^{k} X_j$, $V_k^2 = \sum_{j=1}^{k} \mathrm{E}[X_j^2 | X_1, \ldots, X_{j-1}] = A_k^2$, *and* $u_k^2 = \mathrm{E}[U_k^2] = \mathrm{E}[V_k^2] = b_k^2$. *Note that* $\mathrm{E}[X_j] = 0$, *and that* X_j *by (A1) have uniformly bounded second moments, and hence that* $\max_{j \le k} \mathrm{E}[X_j^2]$ *are uniformly bounded. Since* $u_k \to \infty$ *as* $k \to \infty$, *we therefore have that* $\max_{j \le k} u_k^{-1} X_j \xrightarrow{L_2} 0$, *which implies*

$$\max_{j \le k} u_k^{-1} X_j \xrightarrow{p} 0. \tag{S.5}$$

This in turn is equivalent to the weaker Lindeberg-Feller condition

$$u_k^{-2} \sum_{j=1}^{k} X_j^2 I(|X_j| \ge \varepsilon u_k) \xrightarrow{p} 0 \quad \text{for all } \varepsilon > 0, \tag{S.6}$$

since $P(\max_{j \le k} u_k^{-1} X_j > \varepsilon) = P(\sum_{j=1}^{k} u_k^{-2} X_j^2 I(|X_j| > \varepsilon u_k) > \varepsilon^2)$. *But*

$$\mathrm{E}\left[\left|u_k^{-2} \sum_{j=1}^{k} X_j^2 I(|X_j| \ge \varepsilon u_k)\right|\right] \le u_k^{-2} \mathrm{E}\left[\sum_{j=1}^{k} X_j^2\right] = 1 \quad \text{for all } k. \tag{S.7}$$

By the dominated convergence theorem, (S.6) *and* (S.7) *implies the Lindeberg-Feller condition* (S.1)*, which together with (A3) according to Lemma 1 gives*

$$U_k/u_k \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as } k \to \infty.$$

The first result (S.2) *now follows by noting that*

$$U_k/u_k = \frac{\sum_{j=1}^k n_j(\hat{t}_{y,j} - t_y)}{b_k} = \frac{m_k^{-1}\sum_{j=1}^k n_j(\hat{t}_{y,j} - t_y)}{b_k/m_k} = \frac{\hat{t}_y^{(k)} - t_y}{b_k/m_k}.$$

For (S.3)*, we first note that* $\hat{t}_{y,k} = O_p(1)$ *and* $\hat{t}_y^{(k)} = t_y + O_p(b_k/m_k)$*. Hence*

$$\begin{aligned} b_k^{-2}\sum_{j=1}^k n_j^2\left(\hat{t}_{y,j} - \hat{t}_y^{(k)}\right)^2 &= b_k^{-2}\sum_{j=1}^k n_j^2\left(\hat{t}_{y,j} - t_y + O_p(b_k/m_k)\right)^2 \\ &= b_k^{-2}\sum_{j=1}^k n_j^2\left(\hat{t}_{y,j} - t_y\right)^2 + O_p(b_k^{-1}). \end{aligned}$$

Next,

$$\begin{aligned} \mathrm{E}\left|b_k^{-2}\sum_{j=1}^k n_j^2\left(\hat{t}_{y,j} - \hat{t}_y^{(k)}\right)^2 - 1\right| &= \mathrm{E}\left|b_k^{-2}\sum_{j=1}^k n_j^2\left(\hat{t}_{y,j} - t_y\right)^2 + O_p(b_k^{-1}) - 1\right| \\ &\le \mathrm{E}\left|b_k^{-2}\sum_{j=1}^k n_l^2\left(\hat{t}_{y,j} - t_y\right)^2 - 1\right| + \mathrm{E}[O_p(b_k^{-1})]. \end{aligned}$$

Note now that (A3) is equivalent to $\lim_{k\to\infty}\mathrm{E}|A_k^2 b_k^{-2} - 1|$ *(Brown, 1971, Lemma 1), which together with* (S.5) *and the Lindeberg-Feller condition* (S.1) *implies that the first term vanishes as* $k \to \infty$ *(Hall and Heyde, 1980, Theorem 3.5). As does the second term, since* $\hat{t}_{y,j}$ *have uniformly bounded second moments. Hence,* (S.3) *now follows since convergence in mean implies convergence in probability.*

For the last result (S.4)*, we note that* $\mathrm{E}[\sum_{j=1}^k n_j\hat{\sigma}_j^2] = b_k^2$*, and*

$$\mathrm{Var}\left(b_k^{-2}\sum_{j=1}^k n_j\hat{\sigma}_j^2\right) = \frac{\mathrm{Var}(\sum_{j=1}^k n_j\hat{\sigma}_j^2)}{b_k^2}\frac{1}{b_k^2} \to 0 \quad \text{as } k \to \infty,$$

since the first factor by (A4) is bounded and the second goes to zero as $k \to \infty$*. This completes the proof.* □

Proof *of Corollary S1*

The result follows immediately from Proposition S1 by application of the Cramér-Wold theorem and the Delta method (see, e.g., Sen and Singer, 1993, Theorem 3.2.4 and 3.4.5) .

A.2 Optimality results

We present in this section proofs of the optimality results of Proposition 1–2 and Corollary 1–2. First, two lemmas are presented.

Lemma 2

Let $\boldsymbol{S} = (S_1, \ldots, S_N) \sim \text{Multinomial}(n, \boldsymbol{\pi})$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$, $\boldsymbol{\mu} := \mathrm{E}[\boldsymbol{S}] = (\mu_1, \ldots, \mu_N)$. *Let*

$$\hat{\boldsymbol{t}}_{\boldsymbol{y}} = \sum_{i=1}^{N} S_i w_i \boldsymbol{y}_i, \quad w_i := \mathrm{E}[S_i]^{-1} = (n\pi_i)^{-1}.$$

Then the covariance matrix of $\hat{\boldsymbol{t}}_{\boldsymbol{y}}$ *is given by*

$$\mathbf{Cov}(\hat{\boldsymbol{t}}_{\boldsymbol{y}}) = \frac{1}{n}\left(\sum_{i=1}^{N} \frac{\boldsymbol{y}_i \boldsymbol{y}_i^T}{\pi_i} - \sum_{i,j=1}^{N} \boldsymbol{y}_i \boldsymbol{y}_j^T\right).$$

Lemma 3

Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ *and consider the function*

$$f(\boldsymbol{\pi}) = \sum_{i=1}^{N} \frac{c_i^2}{\pi_i}$$

for some coefficients $c_i > 0$. *Subject to the constraints*

$$\sum_{i=1}^{N} \pi_i = 1, \quad \pi_i > 0,$$

$f(\boldsymbol{\pi})$ *is minimized by* $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_N^*)$ *with*

$$\pi_i^* = \frac{c_i}{\sum_{j=1}^{N} c_j}, \quad i = 1, \ldots, N.$$

Proof *of Lemma 2*

By properties of the multinomial distribution, we have that

$$\mu_i := \mathrm{E}[S_i] = n\pi_i, \quad \mathrm{Var}(S_i) = n\pi_i(1-\pi_i), \quad \mathrm{Cov}(S_i, S_j) = -n\pi_i\pi_j,$$

for $i, j = 1, \ldots, N, i \neq j$. *Hence,*

$$\mathbf{Cov}(\hat{\boldsymbol{t}}_{\boldsymbol{y}}) = \mathbf{Cov}\left(\sum_{i=1}^{N} S_i w_i \boldsymbol{y}_i\right) = \left(\sum_{i=1}^{N} \frac{n\pi_i(1-\pi_i)}{n^2\pi_i^2} \boldsymbol{y}_i \boldsymbol{y}_i^T - \sum_{\substack{i,j=1 \\ i\neq j}} \frac{n\pi_i\pi_j}{n^2\pi_i\pi_j} \boldsymbol{y}_i \boldsymbol{y}_j^T\right)$$

$$= \frac{1}{n}\left(\sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} \boldsymbol{y}_i \boldsymbol{y}_i^T - \sum_{\substack{i,j=1 \\ i\neq j}}^{N} \boldsymbol{y}_i \boldsymbol{y}_j^T\right) = \frac{1}{n}\left(\sum_{i=1}^{N} \frac{\boldsymbol{y}_i \boldsymbol{y}_i^T}{\pi_i} - \sum_{i,j=1}^{N} \boldsymbol{y}_i \boldsymbol{y}_j^T\right).$$

□

Proof *of Lemma 3*
Using the method of Lagrange multipliers (Boyd and Vandenberghe, 2004, Chapter 5), we introduce the auxiliary function

$$\Lambda(\boldsymbol{\pi}, \lambda) = f(\boldsymbol{\pi}) + \lambda g(\boldsymbol{\pi}), \quad g(\boldsymbol{\pi}) = \sum_{i=1}^{N} \pi_i - 1 \ .$$

Critical points of the Lagrangian are found by solving the equation system

$$\nabla\Lambda(\boldsymbol{\pi}, \lambda) = \mathbf{0} \quad \Leftrightarrow \quad \begin{cases} g(\boldsymbol{\pi}) = 0 \\ -\nabla_{\boldsymbol{\pi}} f(\boldsymbol{\pi}) = \lambda \nabla_{\boldsymbol{\pi}} g(\boldsymbol{\pi}) \end{cases} .$$

Since $\frac{\partial f(\boldsymbol{\pi})}{\partial \pi_i} = -c_i^2/\pi_i^2$ *and* $\frac{\partial g(\boldsymbol{\pi})}{\partial \pi_i} = 1$*, this implies that* $\lambda = c_1^2/\pi_1^2 = \ldots = c_N^2/\pi_N^2$*, and further that* $|\pi_i| \propto |c_i|$*. Since* $c_i > 0$*,* $\pi_i > 0$ *and* $\sum_{i=1}^{N} \pi_i = 1$*, we obtain*

$$\pi_i^* = \frac{c_i}{\sum_{j=1}^{N} c_j}. \tag{S.8}$$

Thus, the point $(\boldsymbol{\pi}^*, \lambda^*)$ *with entries* π_i^* *defined according to* (S.8) *and* $\lambda^* = c_1^2/\pi_1^{*2}$ *is a stationary point of* $\Lambda(\boldsymbol{\pi}, \lambda)$*. Hence,* $\boldsymbol{\pi}^*$ *is a stationary point of* $f(\boldsymbol{\pi})$ *under the specified constraints. Specifically,* $\boldsymbol{\pi}^*$ *is a local minimum. Since we consider a convex function over a convex set,* $\boldsymbol{\pi}^*$ *also is a global minimum. This proves the optimality of* (S.8). □

Proof *of Proposition 1*
We want to minimize the approximate variance $\mathrm{AV}(\hat{\theta}) := \nabla h(\boldsymbol{u})^T \boldsymbol{\Phi}_k \nabla h(\boldsymbol{u})\big|_{\boldsymbol{u}=\boldsymbol{t}_{\boldsymbol{y}}}$ *of the estimator* $\hat{\theta}_k := h(\hat{\boldsymbol{t}}_{\boldsymbol{y},k})$*, where* $\boldsymbol{\Phi}_k = \mathbf{Cov}(\hat{\boldsymbol{t}}_{\boldsymbol{y},k} | \boldsymbol{S}_1, \ldots, \boldsymbol{S}_{k-1})$*. Since* $\boldsymbol{S}_k | \boldsymbol{S}_1, \ldots, \boldsymbol{S}_{k-1} \sim$ Multinomial$(n_k, \boldsymbol{\pi}_k)$*, the approximate variance can according to Lemma 2 be written as*

$$\frac{1}{n_k} \nabla h(\boldsymbol{u})^T \left(\sum_{i=1}^{N} \frac{1}{\pi_{ki}} p_i^2 r_i^2 \boldsymbol{y}_i \boldsymbol{y}_i^T - \sum_{i,j=1}^{N} p_i p_j r_i r_j \boldsymbol{y}_i \boldsymbol{y}_j^T \right) \nabla h(\boldsymbol{u}),$$

evaluated at $\boldsymbol{u} = \boldsymbol{t}_{\boldsymbol{y}}$*. As a function of the sampling scheme* $\boldsymbol{\pi}_k$*, minimizing the above expression is equivalent to minimizing*

$$\sum_{i=1}^{N} \frac{p_i^2 r_i^2 \nabla h(\boldsymbol{u})^T \boldsymbol{y}_i \boldsymbol{y}_i^T h(\boldsymbol{u})}{\pi_{ki}} = \sum_{i=1}^{N} \frac{(p_i r_i \nabla h(\boldsymbol{u})^T \boldsymbol{y}_i)^2}{\pi_{ki}},$$

evaluated at $\boldsymbol{u} = \boldsymbol{t}_{\boldsymbol{y}}$*. The result now follows from Lemma 3 with* $c_i = p_i r_i \big| \nabla h(\boldsymbol{u})^T \boldsymbol{y}_i \big|_{\boldsymbol{u}=\boldsymbol{t}_{\boldsymbol{y}}}$.
□

Proof *of Corollary 1*

a) Follows immediately from Proposition 1 with scalar $\boldsymbol{y}_i = y_i$ *and identity mapping* $h(u) = u$.

b) Take $\boldsymbol{y}_i = (y_i, 1)^T$ *and* $h : \mathbb{R}^2 \to \mathbb{R}$ *as* $h(u_1, u_2) = u_1/u_2$. *Then* $\boldsymbol{t}_{\boldsymbol{y}} = (t_y, t_r)^T$, $\theta = h(\boldsymbol{t}_{\boldsymbol{y}}) = t_y/t_r$, $\hat{\boldsymbol{t}}_{\boldsymbol{y}} = (\hat{t}_y, \hat{t}_r)^T$, *and* $\hat{\theta} = h(\hat{\boldsymbol{t}}_{\boldsymbol{y}}) = \hat{t}_y/\hat{t}_r$. *Since* $\nabla h(u_1, u_2) = (1/u_2, -u_1/u_2^2)^T$, *we have that*

$$\left|\nabla h(\boldsymbol{u})^T \boldsymbol{y}_i\right|_{\boldsymbol{u}=\boldsymbol{t}_{\boldsymbol{y}}} = \left|(1/t_r, -t_y/t_r^2) \cdot \binom{y_i}{1}\right| = t_r^{-1}|y_i - t_y/t_r|.$$

By the result of in Proposition 1, we obtain as optimal sampling probabilities

$$\pi^*_{ki} = \frac{t_r^{-1} p_i r_i |y_i - t_y/t_r|}{\sum_{j=1}^N t_r^{-1} p_j r_j |y_j - t_y/t_r|} = \frac{p_i r_i |y_i - t_y/t_r|}{\sum_{j=1}^N p_j r_j |y_j - t_y/t_r|} = \frac{c_i}{\sum_{j=1}^N c_j},$$

with $c_i = p_i r_i v_i$ *and* $v_i = |y_i - t_y/t_r|$. □

Proof *of Proposition 2*

In analogy with the proof of Proposition 1, minimizing the anticipated approximate variance of the estimator $\hat{\theta}_k := h(\hat{\boldsymbol{t}}_{\boldsymbol{y},k})$ *is equivalent to minimizing*

$$\mathrm{E}_{\boldsymbol{R},\boldsymbol{Y}}\left[\sum_{i=1}^N \frac{(p_i R_i \nabla h(\boldsymbol{u})^T \boldsymbol{Y}_i)^2}{\pi_{ki}}\right] = \sum_{i=1}^N \frac{p_i^2}{\pi_{ki}} \mathrm{E}_{R_i,\boldsymbol{Y}_i}\left[(R_i \nabla h(\boldsymbol{u})^T \boldsymbol{Y}_i)^2\right]$$

evaluated at $\boldsymbol{u} = \boldsymbol{t}_{\boldsymbol{y}}$. *By the law of total expectation we further have that*

$$\mathrm{E}_{R_i,\boldsymbol{Y}_i}\left[(R_i \nabla h(\boldsymbol{u})^T \boldsymbol{Y}_i)^2\right] = \mathrm{E}_{R_i}\left[R_i^2 \mathrm{E}_{\boldsymbol{Y}_i}[(h(\boldsymbol{u})^T \boldsymbol{Y}_i)^2 | R_i]\right] = \hat{r}_i \mathrm{E}_{\boldsymbol{Y}_i}[(h(\boldsymbol{u})^T \boldsymbol{Y}_i)^2 | R_i = 1],$$

where the second equality follows since $R_i \sim \text{Bernoulli}(\hat{r}_i)$. *Using the equality* $\mathrm{E}[X^2] = \mathrm{E}[X]^2 + \mathrm{Var}(X)$, *we next have that*

$$\begin{aligned}
\mathrm{E}\left[(\nabla h(\boldsymbol{u})^T \boldsymbol{Y}_i)^2 | R_i = 1\right] &= \mathrm{E}[\nabla h(\boldsymbol{u})^T \boldsymbol{Y}_i | R_i = 1]^2 + \mathrm{Var}\left(\nabla h(\boldsymbol{u})^T \boldsymbol{Y}_i | R_i = 1\right) \\
&= (\nabla h(\boldsymbol{u})^T \mathrm{E}[\boldsymbol{Y}_i | R_i = 1])^2 + \nabla h(\boldsymbol{u})^T \mathbf{Cov}\left(\boldsymbol{Y}_i | R_i = 1\right) \nabla h(\boldsymbol{u}) \\
&= (\nabla h(\boldsymbol{u})^T \hat{\boldsymbol{y}}_i)^2 + \nabla h(\boldsymbol{u})^T \boldsymbol{\Sigma}_i \nabla h(\boldsymbol{u}).
\end{aligned}$$

Combining these results, we have that

$$\mathrm{E}_{\boldsymbol{R},\boldsymbol{Y}}\left[\sum_{i \in \mathcal{D}} \frac{(p_i^2 R_i \nabla h(\boldsymbol{u})^T \boldsymbol{Y}_i)^2}{\pi_{ki}}\right] = \sum_{i=1}^N \frac{p_i^2 v_i}{\pi_{ki}}$$

with $v_i = \hat{r}_i \left[(\nabla h(\boldsymbol{u})^T \hat{\boldsymbol{y}}_i)^2 + \nabla h(\boldsymbol{u})^T \boldsymbol{\Sigma}_i \nabla h(\boldsymbol{u})\right]\big|_{\boldsymbol{u}=\boldsymbol{t}_{\boldsymbol{y}}}$. *The desired result now follows from Lemma 3.* □

Proof *of Corollary 2*
a) Follows immediately from Proposition 2 applied to the scalar case with $\boldsymbol{Y}_i = Y_i$, $\hat{\boldsymbol{y}}_i = \hat{y}_i$, $\boldsymbol{\Sigma}_i = \sigma_i^2$ *and identity mapping* $h(u) = u$.

b) Take $\boldsymbol{Y}_i = (Y_i, 1)^T$, $\hat{\boldsymbol{y}}_i = (\hat{y}_i, 1)^T$ $\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma_i^2 & 0 \\ 0 & 0 \end{pmatrix}$, *and* $h : \mathbb{R}^2 \to \mathbb{R}$ *as* $h(u_1, u_2) = u_1/u_2$. *The result follows by insertion in Proposition 2, in analogy with the proof of Corollary 1 b).* □

B Implementation of sampling methods

Additional details on the implementation of the sampling methods in the empirical evaluation are provided below.

General implementation details. All sampling methods were implemented in an iterative fashion according to Algorithm 1, and differed only in the calculation of the sampling schemes (steps 2–7 in Algorithm 1). The learning and optimization steps were only performed for active sampling. All methods were implemented using multinomial sampling, with replacement within and between iterations. Although this introduces a slight disadvantage for the reference methods, which could also be implemented without replacement, the resulting loss of efficiency is negligible since the overall sampling fraction is small.

Importance sampling methods. Two different importance sampling schemes were considered: density importance sampling and severity importance sampling. With density importance sampling, the sampling probabilities were selected proportional to the prior observation weights p_i, since instances with large observation weights by design of the scenario generation framework have a larger contribution to estimation. Since our aim is safety benefit evaluation of an advanced driver assistance system compared to some baseline driving scenario, and the potential safety benefit increases with impact severity, one may expect that oversampling of high-severity instances will lead to further variance reduction in the safety benefit evaluation. We therefore also included severity importance sampling, which attempts to oversample high-severity instances by assigning sampling probabilities proportional to $p_i \times o_i \times d_i \times m_i$, where p_i is the prior observation weight of instance i, o_i is the corresponding off-road glance duration, d_i is the maximal deceleration, and m_i an *a priori* known maximal possible impact speed of instance i. To account for the difference in scaling between variables, all variables (off-road glance duration, deceleration, maximal impact speed) were transformed a common scale by mapping the values onto the interval

$[0.1, 1]$ before calculating the severity sampling scheme.

Initialization. Active sampling and severity sampling were initialized with a deterministic sample of 44 instances, one per original (real) rear-end crash, at the maximum glance duration 6.6s and minimal deceleration 3.3m/s^2. Hence, the maximal possible impact speed in each scenario could be retrieved. Knowing the maximal impact speed was necessary for severity sampling. It was also assumed to be beneficial for active sampling, as we expected the maximal impact speed to be an important predictor of the case-specific safety benefit response profile. To avoid selection bias, the initial sample did not contribute to estimation and was retained in the sampling frame for selection in future iterations. It was, however, used in the learning step of the active sampling algorithm. With regards to sample size, the initial sample counted as 44 observations when comparing the performance of different sampling methods.

Learning and prediction. For the learning step of the active sampling algorithm, we used random forest regression for continuous outcomes (impact speed reduction) and random forest classification for binary outcomes (crash/no-crash under baseline and countermeasure scenarios) (Breiman, 2001). The random forest method was chosen for the following reasons: i) it is fast and flexible, ii) it is capable of finding non-linear and non-monotonic patterns, as well as interactions between variables, without the need for explicit feature construction, and iii) measures of generalization error and prediction performance are readily available through estimates of residual variance, prediction R-squared and accuracy on hold-out (out-of-bag) data.

Explanatory variables were the input parameters to the simulations (off-road glance duration and maximal deceleration), and case-specific maximal impact speed retrieved from the initialization step described above. Random forest was fitted using 100 trees with variance splitting rule for regression, and gini splitting rule for classification. Other hyper-parameters (minimum node size and number of variables to split upon) were selected with 5-fold cross validation, using a random grid search of minimum node size from 1 to 20 and number of variables to split upon from 1 to 3. All predictions were set equal if the model could not be fitted or produced a prediction R-squared less than 0 on hold-out data, thus resorting to density importance sampling. To reduce computation time, prediction models were updated every 10^{th} new observation up to a sample size of $n = 100$ observations, thereafter every 25^{th} new observation up to a sample size of $n = 500$ observations, thereafter every 50^{th} observation up to a sample size of $n = 1,000$ observations, and so on.