Optimal sampling in unbiased active learning

Citation for the original published paper (version of record):
Imberg, H., Jonasson, J., Axelson-Fisk, M. (2020). Optimal sampling in unbiased active learning. Proceedings of Machine Learning Research, 108: 559-569

(article starts on next page)

Optimal sampling in unbiased active learning

Henrik Imberg **Johan Jonasson** **Marina Axelson-Fisk**

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

Abstract

A common belief in unbiased active learning is that, in order to capture the most informative instances, the sampling probabilities should be proportional to the uncertainty of the class labels. We argue that this produces suboptimal predictions and present sampling schemes for unbiased pool-based active learning that minimise the actual prediction error, and demonstrate a better predictive performance than competing methods on a number of benchmark datasets. In contrast, both probabilistic and deterministic uncertainty sampling performed worse than simple random sampling on some of the datasets.

1 Introduction

Consider a statistical learning problem where we want to estimate a parameter $\boldsymbol{\theta}$ of a statistical model $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ given a random sample $(\boldsymbol{x}_i, y_i), i = 1, \ldots, N$, with the aim of making predictions $\hat{y}_i$ of y_i given input features $\boldsymbol{x}_i$. Suppose further that the features $\boldsymbol{x}_i$ are observed for all i, but that the outcomes y_i are expensive to measure and therefore may be observed only for a subset of size n; this subset is, however, up to us to choose freely.

The setting described above is the set-up of pool-based active learning, an algorithmic framework where a semi-supervised learning algorithm iterates between data collection and model fitting by repeatedly querying the label of new instances from a large pool of unlabelled observations (Settles, 2012). However, this may also be recognised as a problem arising in the field of finite population sampling, where the parameter of interest may be estimated from a subset of elements selected non-uniformly at random by application of suitable sample weighting techniques (Binder, 1983; Skinner, 1989). Indeed, the connection between active learning and finite population sampling has recently been recognised, and several active learning algorithms using finite population sampling methodology have been proposed (Bach, 2007; Beygelzimer et al., 2009; Chu et al., 2011; Ganti and Gray, 2012).

The finite population sampling methodology offers a promising solution to the subset selection problem in active learning, as it allows for oversampling of the most informative instances without compromising unbiasedness. By use of inverse probability weighting, an unbiased estimator of the total loss is obtained, from which consistent estimates of the optimal parameter may be computed. Furthermore, this holds true even under the realistic assumption of model misspecification (Binder, 1983; Skinner, 1989; Pfeffermann, 1993; Yuan and Jennrich, 1998), as opposed to ordinary maximum likelihood estimation or empirical risk minimisation, which produce inconsistent parameter estimates for misspecified models under covariate shift (Shimodaira, 2000; Sugiyama, 2006; Bach, 2007; Sugiyama and Nakajima, 2009). On the other hand, the increase in variance by use of inverse probability weighting may be substantial (Pfeffermann, 1993; Chambers et al., 2012). Thus, the development of unbiased active learning algorithms that yield low variances in the quantities of interest is essential.

For classification problems, it has been suggested that unbiased active learning algorithms with good properties are obtained by use of probabilistic uncertainty sampling, assigning sampling probabilities proportional to the entropy of the label distribution evaluated at the current parameter estimate (Chu et al., 2011; Ganti and Gray, 2012). This seems to be motivated by the heuristic argument that it would capture the most informative instances (Lewis and Gale,

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

1994; Ganti and Gray, 2012), and has further been supported by the assertion that it would minimise the variance of the estimated (logarithmic) loss (Chu et al., 2011). However, we argue that this sampling strategy is suboptimal and reaches none of the desired targets, as it i) disregards the fact that instances vary in influence and informativeness also by their location in the feature space, and ii) targets the variance of the estimated (logarithmic) loss given the expected data, rather than the expected variance given the actual data. Additionally, the idea of minimising the variance of the expected loss generalises poorly to general prediction problems, as it in fact suggests that passive learning, i.e. uniform random sampling, would be optimal for regression problems, clearly an unsatisfactory result.

Contributions Considering a general family of parametric prediction models, we derive an asymptotic expansion for the expected generalisation error and for the mean squared error of the predictions, and consequently present sampling schemes that optimise the performance of the active learning algorithm with respect to these quantities. The resulting sampling schemes depend both on the label uncertainty and on the influence on model fitting through the location of data points in the feature space, and have a close connection to statistical leverage – a commonly used measure of influence in generalised linear regression modelling (McCullagh and Nelder, 1989).

In the next section, we outline the algorithm for unbiased pool-based active learning, introduced by Ganti and Gray (2012), that will be considered in this paper. Our main results are presented in Section 3, where optimal sampling schemes for three different optimality criteria are derived. Specifically, we consider the variance of the estimated loss, the expectation of the total loss of the active learning algorithm, and the mean squared error of the predictions. The suggested sampling procedures are evaluated empirically in Section 4. Proofs of our theoretical results are provided in the online supplementary appendix.

2 Unequal probability sampling in pool-based active learning

Consider a pool $\mathcal{P}$ of N instances, labelled as $i = 1, \ldots, N$. Each member of the pool is associated with an outcome y_i and a feature vector $\boldsymbol{x}_i$, where the features are known for all instances in the pool, but the outcomes may be observed only for a smaller subset. The outcome may be either categorical, as in classification problems, or numeric, as in regression problems. We consider also a statistical model $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$, indexed by a parameter vector $\boldsymbol{\theta}$, and let $\mu(\boldsymbol{x}, \boldsymbol{\theta}) := \mathrm{E}_{\boldsymbol{\theta}}[Y|\boldsymbol{x}]$ denote the conditional mean of the outcome under the model $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$, which we use to make predictions $\hat{y}_i$ of y_i from the features $\boldsymbol{x}_i$. For estimation, we consider a loss function $\ell(y, \boldsymbol{x}, \boldsymbol{\theta})$,[1] describing the loss associated with the prediction derived from the pair $(\boldsymbol{x}, \boldsymbol{\theta})$ when the true outcome is y, and denote by $\ell_i(\boldsymbol{\theta}) = \ell(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})$ the loss associated with an instance $i \in \mathcal{P}$ for a specific parameter value $\boldsymbol{\theta}$. Also, we let

$$\ell_0(\boldsymbol{\theta}) = \sum_{i \in \mathcal{P}} \ell_i(\boldsymbol{\theta}) \tag{1}$$

denote the total loss as a function of $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_0$ the corresponding optimal parameter in the sense that

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}} \ \ell_0(\boldsymbol{\theta}) \, .$$

An active learning algorithm sequentially samples new training examples from the pool of available instances, and retrieves the corresponding labels or outcomes y_i. We let $Q_{t,i}$ be the sample inclusion indicator variable taking the value 1 if instance i is selected in iteration t and 0 otherwise, $\pi_{t,i} := \mathrm{P}(Q_{t,i} = 1)$ denote the corresponding inclusion probability, and $\mathcal{L}_t$ the collection of labelled instances up to and including iteration t. In each step, one instance is selected at random according to a Multinomial$(1, \boldsymbol{\pi}_t)$ distribution, where $\boldsymbol{\pi}_t = (\pi_{t,1}, \ldots, \pi_{t,N})$. Typically, the sampling scheme $\boldsymbol{\pi}_t$ employed in the current iteration will depend on the parameter estimate from the previous iteration. We do, however, postpone the discussion on the choice of sampling scheme to Section 3.

After retrieving the label y_i of the selected instance, the model $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ is updated by choosing $\hat{\boldsymbol{\theta}}_t$ as the minimiser of a weighted loss

$$\hat{\ell}_t(\boldsymbol{\theta}) = \sum_{i \in \mathcal{L}_t} w_{t,i} \ell_i(\boldsymbol{\theta}) \tag{2}$$

for some appropriately chosen weights $w_{t,i}$. Following Ganti and Gray (2012), we use the sampling weights

$$w_{t,i} = \frac{1}{t} \sum_{s=1}^{t} \frac{Q_{s,i}}{\pi_{s,i}} \, , \quad i \in \mathcal{P} \, , \tag{3}$$

which can be computed recursively as

$$w_{t,i} = w_{t-1,i} + \frac{1}{t} \left(\frac{Q_{t,i}}{\pi_{t,i}} - w_{t-1,i} \right) ,$$

starting with $w_{0,i} = 0$. An algorithmic description of this active learning procedure is summarised in Algorithm 1.

[1] We note that a loss function commonly is written as a function $\ell(\hat{y}, y)$ of the prediction $\hat{y}$ and the outcome y, but use the notation $\ell(y, \boldsymbol{x}, \boldsymbol{\theta})$ to emphasise the dependence on the data $(y, \boldsymbol{x})$ and parameter $\boldsymbol{\theta}$.

Using survey sampling terminology, we may, with the choice of sampling weights (3), think of the weighted loss (2) as a Horvitz-Thompson estimator (Horvitz and Thompson, 1952) of the total loss (1). As such, the weighted loss (2) is an unbiased estimator of the total loss (1), provided that all sampling probabilities are strictly positive, following from the fact that the sampling weights have expectation equal to 1; see Ganti and Gray (2012) for additional details. Consequently, active learning algorithms with this property are commonly referred to as unbiased active learners, referring to the unbiasedness of the estimated loss. This property in its turn implies that the weighted estimator $\hat{\boldsymbol{\theta}}_t$ is a consistent estimator of the optimal parameter $\boldsymbol{\theta}_0$ under general regularity conditions. Importantly, this holds even if the model $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ would be misspecified and differ from the actual data generating model; see e.g. Binder (1983); Skinner (1989); Pfeffermann (1993); Yuan and Jennrich (1998) and Bach (2007).

Algorithm 1 Sampling-weighted active learning

Start with an empty sample $\mathcal{L}_0$.

Initialise the sampling weights to $w_{0,i} = 0$ for all $i \in \mathcal{P}$.

1: **for** t = 1, 2, ... **do**

2: Compute sampling probabilities $\pi_{t,i} \in (0,1)$ for all $i \in \mathcal{P}$.

3: Select one instance at random from the pool according to $\boldsymbol{\pi}_t$.

4: Query the value of y_i of the selected instance and add the corresponding index to $\mathcal{L}_t$.

5: Update the sampling weights according to

$$w_{t,i} = w_{t-1,i} + \frac{1}{t}\left(\frac{1}{\pi_{t,i}} - w_{t-1,i}\right), \quad i \in \mathcal{L}_t\,.$$

6: Update the model $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ by choosing

$$\hat{\boldsymbol{\theta}}_t = \arg\min_{\boldsymbol{\theta}} \sum_{i \in \mathcal{L}_t} w_{t,i}\ell_i(\boldsymbol{\theta})\,.^{2} \tag{4}$$

7: **end for**

3 Optimal sampling schemes

For the estimation of a simple population characteristic, such as a finite population mean or total, it is a well known fact that the variance of an inverse probability weighted estimator of the corresponding statistic is minimised by assigning sampling probabilities proportional to the size of the characteristic of interest, commonly referred to as PPS sampling (Hansen and Hurwitz, 1943; Horvitz and Thompson, 1952; Särndal et al., 2003). For predictive modelling, however, there is some ambiguity in what quantity that should be targeted, and consequently what measure of 'size' that should be used. In this section, we consider three different targets relevant for predictive modelling, and consequently derive sampling schemes that minimise the variances, expectations or mean squared errors of the corresponding quantities. We further show that the notion of sampling with probability proportional to size in the context of predictive modelling naturally translates into sampling with probability proportional to 'influence'.

Our first proposition relates to Chu et al. (2011) and considers the variance of the estimated loss at a specific parameter value. In the second proposition, we consider the expected generalisation error in terms of the total loss of the active learning algorithm, $\ell_0(\hat{\boldsymbol{\theta}}_t)$, which is similar to Bach (2007). The third proposition is inspired by Schein (2005) and Schein and Ungar (2007), and considers the mean squared error of the predictions derived from $\hat{\boldsymbol{\theta}}_t$. All propositions are formulated in terms of Algorithm 1 but may immediately be generalised to batch sampling using e.g. Poisson sampling or multinomial sampling, as outlined in Imberg (2019), and to e.g. L_2-penalised loss functions, which we will use in Section 4. Proofs are presented in the online supplementary appendix.

In the propositions that follow, we adopt a finite population sampling viewpoint where the outcomes $y_1, \ldots, y_N$ are considered to be unknown but fixed constants, as motivated by the fact that the random process generating the data in the pool is carried out prior to instance selection; the values of y_i are merely waiting to be observed. The actual values of $\boldsymbol{y} := (y_1, \ldots, y_N)$ are, however, in general unknown to us, and we introduce a collection of random variables $Y_1^*, \ldots, Y_N^*$, distributed according to the model $f_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i)$, to account for our uncertainty about the true values of $\boldsymbol{y}$. We use subscript $\boldsymbol{\theta}$ to denote expectations with respect to $\boldsymbol{Y}^* := (Y_1^*, \ldots, Y_N^*)$ under the model $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$, and subscript $\boldsymbol{\pi}$ to denote expectations and variances with respect to $\boldsymbol{Q}_{1:t} := \boldsymbol{Q}_1, \ldots \boldsymbol{Q}_t$, i.e with respect to the sampling mechanism. Following Isaki and Fuller (1982), we define the anticipated variance of a statistic $\hat{T}(\boldsymbol{y}, \boldsymbol{Q}_{1:t}; \boldsymbol{X}, \boldsymbol{\theta}_0)$ as $\mathrm{E}_{\boldsymbol{\theta}}[\mathrm{Var}_{\boldsymbol{\pi}}(\hat{T}(\boldsymbol{Y}^*, \boldsymbol{Q}_{1:t}; \boldsymbol{X}, \boldsymbol{\theta})\big|\boldsymbol{Y}^*)]$, i.e. as the model-based expectation of the variance of $\hat{T}$ under repeated subsampling; a useful quantity for deriving practically implementable sampling schemes with certain optimality properties.[3] The anticipated mean squared error of a statistic $\hat{T}$ is defined analogously. We may now present our first theoretical result:

[2] We note that this step of the algorithm only can be performed if a sufficient number of instances have been selected, and that a penalised weighted loss may be considered for improved performance in small samples.

[3] By a statistic $\hat{T}(\boldsymbol{y}, \boldsymbol{Q}_{1:t}; \boldsymbol{X}, \boldsymbol{\theta}_0)$, we simply mean a random variable $\hat{T}$ that is a function of $\boldsymbol{y}$, $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_t$, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and, possibly, $\boldsymbol{\theta}_0$.

Proposition 1
Consider the anticipated variance of the contribution to the weighted loss (2) *from iteration* t *at the parameter value* $\tilde{\boldsymbol{\theta}}$*:*

$$\mathrm{E}_{\boldsymbol{\theta}}\left[\mathrm{Var}_{\boldsymbol{\pi}}\left(\frac{1}{t}\sum_{i\in\mathcal{P}}\frac{Q_{t,i}}{\pi_{t,i}}\ell(Y_i^*,\boldsymbol{x}_i,\tilde{\boldsymbol{\theta}})\big|\boldsymbol{Y}^*\right)\right]. \quad (5)$$

As a function of $\boldsymbol{\pi}_t$*, the anticipated variance* (5) *is minimised by choosing sampling probabilities according to*

$$\pi_{t,i}\propto\sqrt{\mathrm{E}_{\boldsymbol{\theta}}[\ell(Y_i^*,\boldsymbol{x}_i,\tilde{\boldsymbol{\theta}})^2]}\,, \quad (6)$$

for all $i\in\mathcal{P}$*, normalised so that* $\sum_{i\in\mathcal{P}}\pi_{t,i}=1$*.*

As opposed to Chu et al. (2011), Proposition 1 suggests that the variance of the estimated loss at a specific parameter value is minimised by sampling with probability proportional to the square root of the expected squared loss, and not by sampling with probability proportional to the expected loss. To highlight the difference, the sampling scheme of Proposition 1 aims to minimise the expected variance of the actual data, while sampling with probability proportional to the expected loss would replace the actual data by its expectation. To implement this sampling scheme in practice, one would typically evaluate the loss at the current parameter estimate and compute the expectations involved in the sampling scheme (6) under the current model, i.e. take $\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}_{t-1}$.

We point out that the result of Proposition 1 has its main use when estimation of the total loss is of primary interest, as is the case in e.g. model validation, and that the sampling scheme (6) not necessarily yields an active learning algorithm that produces optimal predictions. Thus, we provide in Proposition 2 an asymptotic expansion of the generalisation error in terms of the total loss $\ell_0(\hat{\boldsymbol{\theta}}_t)$ of the active learning algorithm, and consequently present a sampling scheme that minimises the expectation of this quantity in Corollary 1a. Similarly, we provide in Proposition 3 an asymptotic expansion of the mean squared error of the predictions derived from $\hat{\boldsymbol{\theta}}_t$, and present a sampling scheme that minimises the expectation of this quantity in Corollary 1b. To do so, we first need to introduce some additional notation and assumptions, as detailed below.

First, assume that

A1 the loss function $\ell(y,\boldsymbol{x},\boldsymbol{\theta})$ is twice differentiable with respect to the parameter $\boldsymbol{\theta}$,

and let $\boldsymbol{S}(\boldsymbol{\theta})$ and $\boldsymbol{H}(\boldsymbol{\theta})$ denote the $(p\times 1)$ gradient vector and the $(p\times p)$ Hessian matrix of the total loss $\ell_0(\boldsymbol{\theta})$, respectively, and let $\hat{\boldsymbol{S}}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{H}}(\boldsymbol{\theta})$ be their corresponding weighted estimators, defined in analogy with the weighted loss (2). Assume further that

A2 the distributions of $\sqrt{t}(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_0)$ and $\frac{\sqrt{t}}{N}(\hat{\boldsymbol{S}}(\boldsymbol{\theta})-\boldsymbol{S}(\boldsymbol{\theta}))$ with respect to the subsampling mechanism converge to normal laws with zero mean and non-degenerate covariance matrices as $N\to\infty$, $t\to\infty$ and $N-t\to\infty$,

A3 the limit of $\frac{1}{N}\hat{\boldsymbol{H}}(\boldsymbol{\theta}_0)$ as $N\to\infty$, $t\to\infty$ and $N-t\to\infty$ exists and equals the limit of $\frac{1}{N}\boldsymbol{H}(\boldsymbol{\theta}_0)$, and that the matrix $\boldsymbol{H}(\boldsymbol{\theta}_0)$ has full rank.

For Proposition 3 we also need the following assumption:

A4 the mean function $\mu(\boldsymbol{x},\boldsymbol{\theta}):=\mathrm{E}_{\boldsymbol{\theta}}[Y|\boldsymbol{x}]$ is differentiable with respect to the parameter $\boldsymbol{\theta}$.

We note that assumptions A1 and A4 are immediately fulfilled for a wide range of statistical models, including e.g. generalised linear models (McCullagh and Nelder, 1989). Conditions necessary for assumptions A2 and A3 to hold are given in the literature; see e.g. Binder (1983) and references therein.[4]

Finally, we let $\boldsymbol{X}$ denote the $(N\times p)$ model matrix with rows $\boldsymbol{x}_i^T$, $\boldsymbol{s}_i(\boldsymbol{\theta})=\nabla_{\boldsymbol{\theta}}\ell(y_i,\boldsymbol{x}_i,\boldsymbol{\theta})$ the $(p\times 1)$ gradient vector of the loss pertaining to instance i, $\boldsymbol{M}(\boldsymbol{\theta})$ the $(N\times p)$ matrix with rows $\nabla_{\boldsymbol{\theta}}\mu(\boldsymbol{x}_i,\boldsymbol{\theta})^T$, and $||\boldsymbol{v}||$ the Euclidean norm of a vector $\boldsymbol{v}$, i.e. $||\boldsymbol{v}||=\sqrt{\boldsymbol{v}^T\boldsymbol{v}}$. We are now ready to formulate our main results.

Proposition 2
Let $\boldsymbol{y}:=(y_1,\ldots,y_N)$ *and* $\boldsymbol{\pi}_{1:t}:=(\boldsymbol{\pi}_1,\ldots,\boldsymbol{\pi}_t)$ *be fixed, and assume that A1 - A3 holds. Then, the expected generalisation error* $\mathrm{E}_{\boldsymbol{\pi}}[\ell_0(\hat{\boldsymbol{\theta}}_t)]$ *admits the asymptotic expansion*

$$\mathrm{E}_{\boldsymbol{\pi}}\left[\frac{1}{N}\ell_0(\hat{\boldsymbol{\theta}}_t)\right]=\frac{1}{N}\ell_0(\boldsymbol{\theta}_0)+\frac{1}{2Nt^2}\sum_{s=1}^{t}\sum_{i=1}^{N}\frac{c_i(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\theta}_0)}{\pi_{s,i}}+k+o(t^{-1})\,, \quad (7)$$

where

$$c_i(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\theta})=s_i(\boldsymbol{\theta})^T\boldsymbol{H}(\boldsymbol{\theta})^{-1}s_i(\boldsymbol{\theta})$$

and k *is a constant not depending on* $\boldsymbol{\pi}_{1:t}$*. Moreover, the second term of* (7) *is minimised by choosing* $\boldsymbol{\pi}_{1:t}$ *according to*

$$\pi_{s,i}\propto\sqrt{c_i(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\theta}_0)} \quad (8)$$

[4]Formally, we consider in assumption A2 and A3 a (hypothetical) sequence $\mathcal{P}_1,\mathcal{P}_2,\ldots$ of pools of increasing sizes $N_1,N_2,\ldots$, but leave the dependence on this sequence implicit our somewhat simplified notation.

for all $i \in \mathcal{P}$ *and* $s = 1, \ldots, t$*, normalised so that* $\sum_{i \in \mathcal{P}} \pi_{s,i} = 1$.

Proposition 3
Let $\boldsymbol{y} := (y_1, \ldots, y_N)$ *and* $\boldsymbol{\pi}_{1:t} := (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_t)$ *be fixed, and assume that A1 - A4 holds. Then, the mean squared error of the predictions* $\{\mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_t)\}_{i \in \mathcal{P}}$ *admits the asymptotic expansion*

$$\mathrm{E}_{\boldsymbol{\pi}}\left[\frac{1}{N}\sum_{i=1}^{N}\Big(\mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_t) - \mu(\boldsymbol{x}_i, \boldsymbol{\theta}_0)\Big)^2\right] = \frac{1}{Nt^2}\sum_{s=1}^{t}\sum_{i=1}^{N}\frac{d_i(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta}_0)}{\pi_{s,i}} + k + o(t^{-1}), \quad (9)$$

where

$$d_i(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta}) = ||\boldsymbol{M}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})^{-1}\boldsymbol{s}_i(\boldsymbol{\theta})||^2$$

and k *is a constant not depending on* $\boldsymbol{\pi}_{1:t}$*. Moreover, the first term of* (9) *is minimised by choosing* $\boldsymbol{\pi}_{1:t}$ *according to*

$$\pi_{s,i} \propto \sqrt{d_i(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta}_0)} \quad (10)$$

for all $i \in \mathcal{P}$ *and* $s = 1, \ldots, t$*, normalised so that* $\sum_{i \in \mathcal{P}} \pi_{s,i} = 1$.

We emphasise that the results of Proposition 2 and 3 are based on a finite population sampling viewpoint where the outcomes y_i are considered to be fixed but unknown constants. Consequently, all randomness involved is ascribed to the selection mechanism and the results are free of modelling assumptions in the sense that they do not rely on $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ being the model from which the data was actually generated. On the other hand, these results are rather impractical, since computation of the sampling schemes (8) and (10) would require the outcomes y_i and the optimal parameter $\boldsymbol{\theta}_0$ to be known. In practice, one may thus instead aim to minimise the model-based expectations of these quantities. This amounts to replacing $c_i(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta}_0)$ and $d_i(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta}_0)$ in (8) and (10) by $\mathrm{E}_{\boldsymbol{\theta}}[c_i(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{\theta})]$ and $\mathrm{E}_{\boldsymbol{\theta}}[d_i(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{\theta})]$, where the expectation is taken with respect to $Y_1^*, \ldots, Y_N^*$ under the model $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$, and $\boldsymbol{\theta}$ may be taken as the current parameter estimate $\hat{\boldsymbol{\theta}}_{t-1}$. Such a procedure provides practically implementable approximations to the unknown optimal sampling schemes (8) and (10) that ideally would approach optimal performance.

As an example, we present in Corollary 1 two practically implementable sampling schemes that are optimised to minimise the anticipated generalisation error in terms of the total loss of the active learning algorithm (Corollary 1a), and to minimise the anticipated mean squared error of the predictions (Corollary 1b) for a class of generalised linear models (McCullagh and Nelder, 1989).

Corollary 1
Consider a generalised linear model with canonical link function. Let $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ *be the* $(N \times N)$ *diagonal matrix with entries* $\mathrm{Var}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i)$*, and assume that* $\boldsymbol{H} = \boldsymbol{H}(\boldsymbol{\theta}) \propto \boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{X}$ *has full rank. Then,*

(a) the anticipated asymptotic generalisation error, given by the model-based expectation of $\mathrm{E}_{\boldsymbol{\pi}}\left[\ell_0(\hat{\boldsymbol{\theta}}_t)\right]$*, is minimised by choosing*

$$\pi_{t,i} \propto \sqrt{h_{ii}(\boldsymbol{\theta})}, \quad (11)$$

where $h_{ii}(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i)\boldsymbol{x}_i^T\boldsymbol{H}^{-1}\boldsymbol{x}_i$ *is the (scaled) statistical leverage score pertaining to instance* i*,*

(b) the anticipated asymptotic mean squared error of the predictions $\{\mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_t)\}_{i \in \mathcal{P}}$ *is minimised by choosing*

$$\pi_{t,i} \propto ||\mathrm{SD}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i)\boldsymbol{V}\boldsymbol{X}\boldsymbol{H}^{-1}\boldsymbol{x}_i||, \quad (12)$$

where $\mathrm{SD}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i) = \sqrt{\mathrm{Var}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i)}$.

For a linear regression model with constant error variance, (11) *and* (12) *coincide and simply further to*

$$\pi_{t,i} \propto \sqrt{\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i},$$

where $\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i =: h_{ii}$ *is the statistical leverage score for linear regression.*

As highlighted by Corollary 1, the optimal sampling schemes suggested by the discussion succeeding Proposition 2 and 3 are not simply functions of the label uncertainty alone, but depend also on the location of data points in the feature space and account for additional problem specific information captured by the Hessian of the total loss and the gradients of the individual losses and predictions. Furthermore, we may interpret the quantities $\mathrm{E}_{\boldsymbol{\theta}}[c_i(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{\theta})]$ and $\mathrm{E}_{\boldsymbol{\theta}}[d_i(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{\theta})]$ used in the practical approximations of the (unknown) optimal sampling schemes (8) and (10) as measures of influence, measuring the anticipated influence of individual data points on the total loss of the active learning algorithm and on the resulting predictions. Indeed, Corollary 1 shows that the suggested sampling schemes have a close connection to statistical leverage – an influence measure commonly used in generalised linear regression modelling (Pregibon, 1981; McCullagh and Nelder, 1989; Rawlings et al., 1998).

To conclude this section, an illustration of the sampling scheme used in Chu et al. (2011) and Ganti and Gray (2012) is provided in Figure 1, along with the sampling schemes suggested by Proposition 1 and Corollary 1.

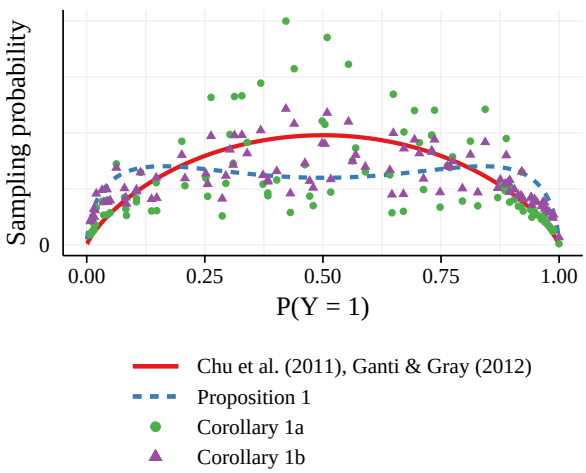

Figure 1: Sampling probability and label uncertainty according to probabilistic uncertainty sampling (Chu et al., 2011; Ganti and Gray, 2012), and according to Proposition 1 and Corollary 1 for a binary classification problem using logistic regression on a simulated dataset with 5 features.

4 Empirical evaluation[5]

4.1 Methods

We evaluated the empirical performance of the proposed sampling schemes for unbiased active learning on binary classification problems on six different data sets: the Abalone dataset, the Statlog Australian Credit Approval and German Credit Data datasets, the Red and White Wine Quality datasets (Cortez et al., 2009), and a dataset derived from the DNA sequence of the E. coli bacteria. The first five datasets were retrieved from the UCI Machine learning repository (Dua and Graff, 2019). For the wine quality datasets, a cut-off at > 6 was used to create a binary outcome from the wine quality score.

The E. coli dataset was constructed by extracting the DNA sequences of all coding sequences (i.e. genes) and all non-coding sequences located between genes. As predictors, we used the relative frequency of two-letter "words" AA, AC, AG, AT, CA, CC, ..., TT made from the "alphabet" A, C, G and T. Thus, there are 16 such word frequencies, used to predict whether a sequence is coding or non-coding. The genomic data, including the complete DNA sequence with annotations, was retrieved from GenBank (2017) (Clark et al., 2016). A summary of the characteristics of the six datasets considered is presented in Table S1 in Appendix C in the online supplement.

Active learning was performed using logistic regression for prediction and implemented according to Algorithm 1, with the following modifications. First, an L_2 (ridge) penalty term was incorporated in the weighted loss function (2) in the estimation step (4), to enable estimation and improve performance in small samples. The penalty parameter was chosen using leave-one-out cross-validation for samples of sizes less than or equal to 50, and 10-fold cross-validation otherwise. To reduce computation time, instances were sampled in batches rather than one at a time, starting with a simple random sample of 25 instances. In each iteration, 25 new instances were sampled from the pool of unlabelled instances using multinomial sampling, thus avoiding unnecessary re-queries on already labelled instances. To retain unbiasedness, labelled instances were re-queried with probability 1 in all succeeding iterations. This comes with no additional cost since the labels of those instances are known already.

The following sample selection procedures were considered: sampling with probabilities computed according to Proposition 1; sampling with probabilities computed according to Corollary 1a (hereafter also referred to as leverage sampling) and Corollary 1b, replacing the Hessian of the ordinary loss by the Hessian of the corresponding L_2-penalised loss; and probabilistic uncertainty sampling, assigning sampling probabilities proportional to the entropy of the label distribution (Chu et al., 2011; Ganti and Gray, 2012). For reference, we also implemented deterministic uncertainty sampling (Lewis and Gale, 1994), querying the labels of the most uncertain unlabelled instances based on the entropy of the label distribution, and passive learning, i.e. uniform random sampling.

Predictive performance was evaluated as follows. The generalisation error was measured by the total loss of the active learning algorithm, i.e. by the negative log-likelihood of the predicted class probabilities on the entire dataset, and the performance on binary classification was measured by the misclassification rate on the entire dataset and by the proportion of correctly classified minority examples, using 50% probability cut-off. The discriminative ability was further assessed by the area under the receiver operating characteristic curve (AUC), which compared to the misclassification rate has the advantage of not being dependent on any particular probability cut-off. The accuracy of the predicted class probabilities was further assessed by the root mean squared error (RMSE) of the predictions, as compared to the predictions obtained when using the entire dataset for training.

[5] A preliminary version of these experiments has previously been presented as a manuscript in Imberg (2019). In the preparation of this paper, an error in the implementation of deterministic uncertainty sampling has been detected and corrected, and the results and discussion have been updated accordingly.

To further assess the bias or unbiasedness of the predictions, we also evaluated the calibration of the predicted class probabilities to the observed outcomes by calculating the calibration slope, as described by Steyerberg and Vergouwe (2014), and by calculating the ratio of observed vs. predicted (expected) number of minority examples. A calibration slope > 1 corresponds to conservative predictions that are shrunk towards the overall mean, and < 1 to overfitting in the sense that the predicted class probabilities are too extreme: low predictions too low and high predictions too high.

The active learning procedure was repeated 10 000 times, and the average or median of the performance metrics was computed. Finally, we estimated the sample sizes required for active learning to achieve equal performance as simple random sampling of $n = 50, \ldots, 250$ instances in terms of AUC, misclassification rate, negative log-likelihood and RMSE, i.e. estimating the sample size reduction that could be achieved by use of active rather than passive learning.

All computations were performed in R v. 3.5.2 (R Core Team, 2018), using the doParallel package v. 1.0.15 (Microsoft Corporation and Weston, 2018) for parallel computing and the glmnet package v. 2.0-18 (Friedman et al., 2010) for fitting ridge logistic regression models. The full code used for the experiments is provided at `https://github.com/imbhe/OSiUAL`.

4.2 Results

The predictive performance in terms of the misclassification rate and negative log-likelihood of the predicted class probabilities for the various active learning algorithms is presented in Figure 2 on the Abalone, E. coli and Red Wine datasets, and in Figure S1 - S7 in Appendix C in the online supplement for all datasets and performance metrics. The impact of active learning on label complexity in terms of the number of instances that need to be queried in order to achieve equal performance as passive learning with a given sample size is presented for all benchmark datasets in Figure S8 - S10 and Table S2 in Appendix C in the online supplement.

As expected, leverage sampling (Corollary 1a) overall achieved the best performance in terms of the negative log-likelihood of the predictions (Figure 2 and S4, Table S2). Similarly, the sampling scheme optimised towards the mean squared error of the predictions (Corollary 1b) produced, together with leverage sampling, the most accurate predictions in terms of the RMSE of the predicted class probabilities (Figure S5, Table S2). Indeed, the two sampling schemes of Corollary 1 had almost identical performance. Both performed better than passive learning on all datasets and performance metrics, and reduced the label complexity for classification by up to 23% (median 8%) compared to passive learning (Table S2). In contrast, the two probabilistic sampling schemes that were determined by label uncertainty alone (Proposition 1, probabilistic uncertainty sampling) both performed worse than passive learning with respect to essentially all performance metrics on four of the datasets.

The predictions obtained by the probabilistic sampling procedures were fairly well-calibrated on most of the datasets, with a slight bias in the predicted probabilities towards the overall mean, as expected by the use of a penalised loss function (Figure S6 and S7). Deterministic uncertainty sampling generated the best classification results (Figure 2, Figure S1 - S2), but performed worse than passive learning with respect to the AUC, RMSE and negative log-likelihood of the predictions on four of the datasets (Figure 2, Figure S3 - S4), and produced poorly calibrated predictions with a severe bias towards the majority class in two of the examples (Figure S6) and overfitted class probability estimates on the majority of the examples (Figure S7).

5 Conclusion

We have studied the impact of the choice of sampling scheme on the statistical properties of unbiased active learning algorithms, conducted an asymptotic analysis of the generalisation error and prediction error, and derived sampling schemes that minimise the expectations of these quantities. We have shown that optimal predictive performance is achieved by oversampling influential instances and high-leverage data points, and that uncertain instances not necessarily are informative ones. Thus, our results stand in contrast to existing algorithms that suggest the use of probabilistic uncertainty sampling (Chu et al., 2011; Ganti and Gray, 2012). Influence-based sampling schemes, on the other hand, have recently been suggested for linear and generalised linear regression modelling in big data applications (Ma et al., 2014; Ma and Sun, 2015; Ma et al., 2015; Wang et al., 2018), as further supported by our findings.

Although our theoretical results are based on asymptotic arguments, we have demonstrated that major improvements in predictive performance may be achieved already at moderate sample sizes. However, the optimality of our theoretical results is somewhat compromised by replacing unknown quantities by their expectations, and by not knowing the optimal parameter. Consequently, only minor improvements were observed on some of the datasets, as compared to passive learning. On the other hand, the suggested sampling

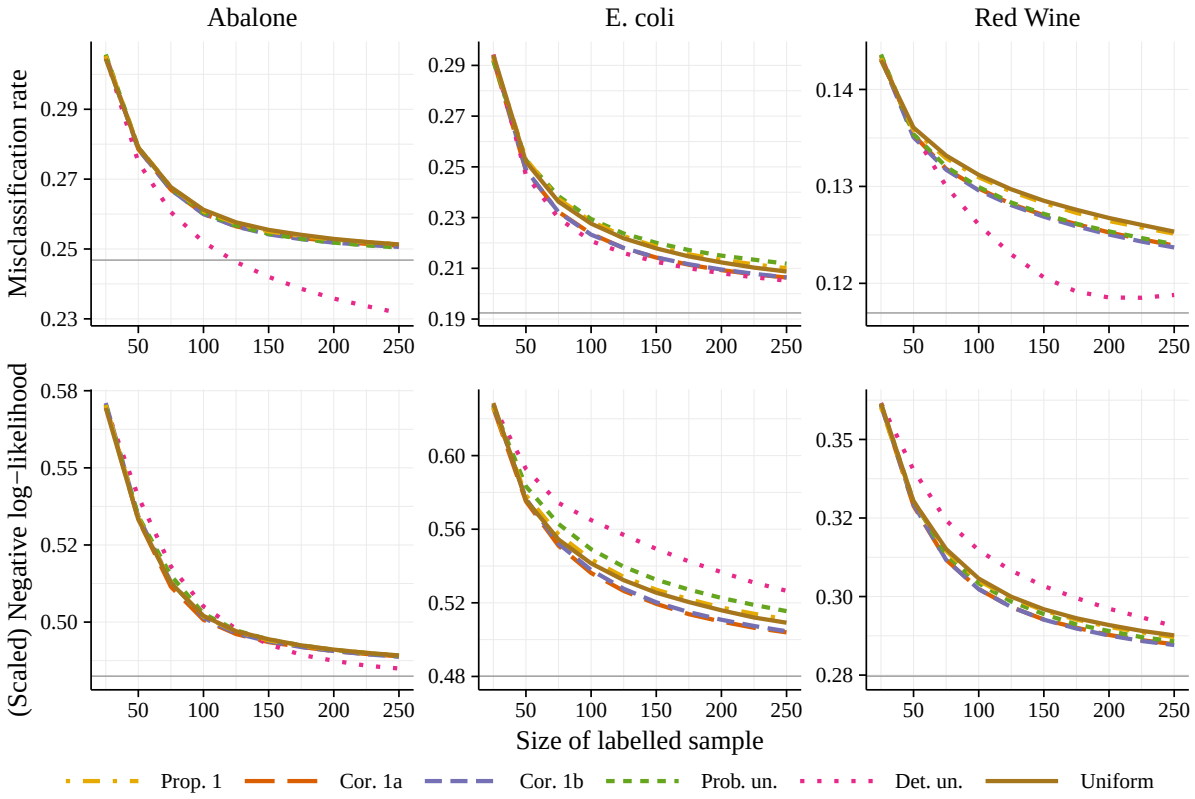

Figure 2: Predictive performance in terms of average misclassification rate and median of the negative log-likelihood (scaled by a factor $1/N$) of the predictions in 10 000 active learning experiments on three benchmark datasets, using sampling schemes optimised to minimise the anticipated variance of the estimated loss (Proposition 1), to minimise the total loss of the active learning algorithm (leverage sampling, Corollary 1a), to minimise the anticipated mean squared error of the predictions (Corollary 1b), probabilistic uncertainty sampling (Chu et al., 2011; Ganti and Gray, 2012), deterministic uncertainty sampling (Lewis and Gale, 1994), and uniform random sampling. The grey solid line shows the performance when using the entire dataset for training.

schemes that were optimised towards predictive performance never performed worse than passive learning, and performed markedly better than passive learning in the majority of the examples.

Somewhat unexpectedly, the best performance in terms of misclassification rate was obtained by deterministic uncertainty sampling, which for some of the datasets even achieved a lower misclassification rate than when using the entire dataset for training. At the same time, deterministic uncertainty sampling performed worse than passive learning with respect to essentially all other performance metrics on the majority of the datasets. Our experiments also revealed a substantial loss of accuracy in the predicted class probabilities, sometimes with a bias towards the majority class and often with an over-optimism in the certainty of the predictions. Despite the positive classification results, our study therefore raises major concerns about the use and applicability of deterministic sample selection procedures, in particular in applications where unbiased estimates of class probabilities are required.

To conclude, our study reveals serious problems with the use of seemingly well-performing deterministic selection procedures in active learning, and demonstrates the benefits of unbiased active learning through unequal probability sampling with inverse probability weighting. Our study further provides a unified framework for optimal sampling in unbiased active learning that is applicable to both regression and classification problems, and suggests that instance selection in unbiased active learning primarily should be influence-driven rather than uncertainty-driven. Even though our theoretical results are limited to regular parametric models and smooth loss functions, we postulate that the same conclusion holds generally, although the notion of influence certainly is model dependent, and studies of sampling strategies for non-regular problems and non-parametric methods are encouraged.

Acknowledgements

We thank Olle Nerman for valuable contributions in the early stages of this work.

References

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

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

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

Chambers, R. L., Steel, D. G., Wang, S., and Welsh, A. (2012). Alternative likelihood-based methods for sample survey data. In *Maximum Likelihood Estimation for Sample Surveys*, chapter 3, pp. 55–88. Chapman & Hall, Boca Raton.

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*.

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2016). GenBank. *Nucleic Acids Research*, 44(D1):D67–D72.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47:547–553.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer, New York.

Dua, D. and Graff, C. (2019). *UCI Machine Learning Repository*. University of California, Irvine. URL `http://archive.ics.uci.edu/ml`.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22. URL `https://www.jstatsoft.org/article/view/v033i01`.

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

GenBank (2017). Escherichia coli O157:H7 str. Sakai DNA, complete genome. Accession No. NC_002695. URL `https://www.ncbi.nlm.nih.gov/nuccore/15829254`.

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. (2019). *Unequal Probability Sampling in Active Learning and Traffic Safety*. Licentiate thesis, Chalmers University of Technology, Gothenburg.

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.

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

Ma, P., Mahoney, M. W., and Yu, B. (2014). A statistical perspective on algorithmic leveraging. In *Proceedings of the 31 st International Conference on Machine Learning*.

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. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(1):70–76.

Mathai, A. and Provost, S. (1992). *Quadratic Forms in Random Variables*. Taylor & Francis, New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, Boca Raton.

Microsoft Corporation and Weston, S. (2018). doParallel: Foreach parallel adaptor for the 'parallel' Package. URL `https://cran.r-project.org/package=doParallel`.

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

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

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. URL `https://www.r-project.org/`.

Rawlings, J. O., Pantula, S. G., and Dickey, D. A. (1998). *Applied Regression Analysis: A Research Tool*. Springer, New York.

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

Schein, A. I. (2005). *Active Learning for Logistic Regression*. PhD thesis, University of Pennsylvania, Philadelphia.

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

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

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

Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In Skinner, C. J., Holt, D., and Smith, T. M. F. (Eds.), *Analysis of Complex Surveys*, pp. 80–87. Wiley, Chichester.

Steyerberg, E. W. and Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal (2014)*, 35:1925–1931.

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

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

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

Yuan, K. H. and Jennrich, R. I. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis*, 65:245–260.