The Canadian Journal of Statistics
Vol. 46, No. 1, 2018, Pages 104–122
La revue canadienne de statistique

Minimax robust active learning for approximately specified regression models

Rui NIE[1], Douglas P. WIENS[2]* and Zhichun ZHAI[2]

[1]*Division of Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital/McGill University, Montreal, QC, Canada, H3T 1E2*
[2]*Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1*

Key words and phrases: Approximate linearity; bias; experimental design; influence; leverage; minimax; model uncertainty; passive learning; sampling variation.

MSC 2010: Primary 62G35, 68T05; secondary 62D99, 62K99

Abstract: We address problems of model misspecification in active learning. We suppose that an investigator will sample training input points (predictors) from a subpopulation with a chosen distribution, possibly different from that generating the underlying whole population. This is in particular justified when full knowledge of the predictors is easily acquired, but determining the responses is expensive. Having sampled the responses the investigator will estimate a, possibly incorrectly specified, regression function and then predict the responses at all remaining values of the predictors. We derive functions $r(\boldsymbol{x})$ of the predictors $\boldsymbol{x}$, and carry out probability weighted sampling with weights proportional to $r(\boldsymbol{x})$. The functions $r(\cdot)$ are asymptotically minimax robust against the losses incurred by random measurement error in the responses, sampling variation in the inputs, and biases resulting from the model misspecification. In our applications the values of $r(\cdot)$ are functions of the diagonal elements of the "hat" matrix which features in a regression on the entire population; this yields an interpretation of sampling the "most influential" part of the population. Applications on simulated and benchmark data sets demonstrate the strong gains to be achieved in this manner, relative to passive learning and to previously proposed methods of active learning. We go on to illustrate the methods in the context of a case study relating ice thickness and snow depth at various locations in Canada, using a "population" of about 50,000 observations made available by Statistics Canada. *The Canadian Journal of Statistics* 46: 104–122; 2018 © 2017 Statistical Society of Canada

Résumé: Les auteurs étudient les problèmes liés à la mauvaise spécification d'un modèle en apprentissage actif. Ils supposent qu'un chercheur obtient ses données d'apprentissage en échantillonnant une sous-population selon une distribution choisie qui peut être différente de celle sous-jacente à la population entière. Cette approche est particulièrement pertinente lorsqu'il est facile d'obtenir la valeur des prédicteurs, mais dispendieux d'obtenir la variable réponse. Après avoir obtenu des variables réponses, le chercheur estime une fonction de régression, possiblement mal spécifiée, et s'en sert pour prédire les réponses à toutes les valeurs de prédicteurs restantes. Les auteurs suggèrent une fonction $r(\boldsymbol{x})$ des prédicteurs $\boldsymbol{x}$ afin de procéder à un échantillonnage pondéré avec des poids proportionnels à $r(\boldsymbol{x})$. Les fonctions $r(\cdot)$ sont asymptotiquement minimax et robustes aux erreurs aléatoires de mesure dans la réponse, aux variations échantillonnales dans les entrées et aux biais résultant d'une mauvaise spécification. Les valeurs de la fonction $r(\cdot)$ utilisée par les auteurs sont basées sur les éléments diagonaux de la matrice chapeau associée à la régression sur la population entière. Ainsi, leur méthode s'interprète comme l'échantillonnage de la partie de la population ayant la plus grande influence. Les auteurs présentent des exemples sur des données réelles et simulées afin de démontrer les gains importants qui peuvent être

Additional supporting information may be found in the online version of this article at the publisher's website.
* *Author to whom correspondence may be addressed.*
E-mail: doug.wiens@ualberta.ca

obtenus en comparaison de l'apprentissage passif ou des méthodes existantes d'apprentissage actif. Ils illustrent leur méthode dans le contexte d'une étude de cas sur l'épaisseur de la glace et de la neige à différents endroits au Canada en utilisant une population de 50 000 observations rendues disponibles par Statistique Canada. *La revue canadienne de statistique* 46: 104–122; 2018 © 2017 Société statistique du Canada

1. INTRODUCTION AND SUMMARY

In this article we consider a variety of sampling paradigms for "active learning," in the face of possible model misspecification. We initially suppose that training input points $\boldsymbol{x} \in \mathbb{R}^k$, arising in a population according to an unknown "test" density $q(\boldsymbol{x})$, will be sampled from a subpopulation with "training" density $p(\boldsymbol{x})$. In "passive learning" $p = q$ and random sampling is carried out from the whole population; our intention is instead that p—more precisely, $r = p/q$—will be selected by the user in accordance with one of several optimality criteria.

The criteria we employ are all aimed at achieving robustness against "model uncertainty." Specifically, we suppose that, having chosen training inputs $\{\boldsymbol{x}_i \mid i = 1, \ldots, n\}$ an investigator will observe, with measurement error, training outputs $\{Y_i\}$. He will then fit a linear parametric response

$$E\left[Y|\boldsymbol{x}\right] = \boldsymbol{f}'\left(\boldsymbol{x}\right)\boldsymbol{\theta} \tag{1}$$

to the training data, with the intention of estimating the parameters, and predicting responses at unsampled points. A motivation for the search for optimality is that the information about the response function is to be acquired through sampling a relatively small number of responses; this is especially relevant if full knowledge of the population of inputs is easily acquired, but sampling responses is expensive.

The investigator is aware that the fitted model is perhaps only a tentative approximation, and seeks protection against the increased estimation and prediction errors arising from alternatives

$$E\left[Y|\boldsymbol{x}\right] = \boldsymbol{f}'\left(\boldsymbol{x}\right)\boldsymbol{\theta} + \psi\left(\boldsymbol{x}\right) \tag{2}$$

for some function ψ satisfying conditions given in Section 2. These conditions ensure that the parameter $\boldsymbol{\theta}$ is identifiable, and that the biasing effect of the model misspecification is bounded.

The literature on the robustness issues in active learning appears to be weighted towards "classification" and "discrimination" problems. Liu & Ziebart (2014) propose methods of "robust estimation" in order to deal with errors arising when the conditional parametric distribution $P_{\boldsymbol{\theta}}\left(Y|\boldsymbol{x}\right)$ is correctly specified but incorrectly estimated; the densities q and p, while possibly different, are fixed. Wen, Yu, & Greiner (2014) consider minimax estimation problems in active learning, with the "max" evaluated over the weighting functions in weighted likelihood estimates that are then optimized. Lanckriet et al. (2002) consider binary classification, seeking (minimax) robustness against improperly specified mean and covariance parameters in the two classes. Kim & Boyd (2008) discuss more general minimax theory with applications in machine learning, signal processing, and finance. For other applications see Cohn, Ghahramani, & Jordan (1996), Scheffer, Decomain, & Wrobel (2001), Tong & Koller (2002), and Tur, Hakkani-Tür, & Schapire (2005). For a systematic review of the active learning literature see Settles (2009).

Somewhat closer to the work presented here is that of Kanamori & Shimodaira (2003), who consider weighted maximum likelihood estimation of an approximate linear response as in (2), after sampling the training points from a "parametric" design density. The weights are chosen so as to minimize the bias, and then the design parameters are optimized. Sugiyama (2006) continues this approach, but uses least squares as the estimation method. The training input density is one of a finite set of densities, chosen empirically on the basis of the performance in preliminary samples. In Sugiyama, Krauledat, & Müller (2007) this input density is given and fixed, and combined with "importance weighted cross validation," in which the importance weighting approach to covariate shift of Shimodaira (2000) is employed.

As pointed out by Settles (2009), active learning for regression problems is essentially what we statisticians call "optimal experimental design." A partial motivation for this article is to investigate the extent to which previously derived methods of "minimax design robustness" against model misspecification, as initially expounded by Huber (1975) and progressing as described in Wiens (2015), can contribute in this context. In design there is a focus on deriving a "design measure," which can be viewed as an empirical distribution on the set of possible design points; the experimenter then chooses inputs dictated deterministically by this distribution. In the mathematical development of the results presented here we first assume—as in Fukumizu (2000), Kanamori & Shimodaira (2003), and Sugiyama (2006)—a density $q(\boldsymbol{x})$ generating the population, with respect to which the focus is on learning the mean response through robustly "sampling" from $p(\boldsymbol{x})$. In all cases p is of the form $p(\boldsymbol{x}) = r(\boldsymbol{x})q(\boldsymbol{x})$, where r does not depend upon q. That the underlying population—the analogue of the design space—has a density q is a feature of active learning which is absent from design theory; however, the analogue of the design measure is seen to be the ratio $r = p/q$, and we find that methods shown to be useful in design can be brought into play to derive optimal choices of this ratio. We propose that the training input points be obtained by probability weighted sampling, with weights proportional to $r(\boldsymbol{x})$; the sampled points then have unconditional density $p(\boldsymbol{x})$. To carry this out it is not necessary that $q(\boldsymbol{x})$ be known or specified.

Cohn, Ghahramani, & Jordan (1996) initiated a study of model-based active learning in regression problems, assuming however that the model being fitted by the investigator was correct. If this assumption fails then the estimates and predictions are biased, and in discussing future work Cohn, Ghahramani, & Jordan (1996) state that "... (the most important problem to be addressed) ... is active bias minimization. ... The variance-minimizing strategy examined here ignores the bias component, which can lead to significant errors when the learner's bias is non-negligible." This mirrors very similar statements used to justify "robustness of design": "... the optimal design in typical situations in which both variance and bias occur is very nearly the same as would be obtained if variance were ignored completely and the experiment designed so as to minimize the bias alone." (Box & Draper 1959, p. 622).

We start by evaluating the integrated, or averaged, (over the possible set χ of inputs) mean squared error (MSE) of the predicted values $\hat{Y}(\boldsymbol{x})$ of outputs. The parameters are estimated by least squares, possibly weighted. This MSE has three components—the variation arising from random measurement error, the (squared) bias incurred in fitting an incorrectly specified response function, and the variation in the bias incurred in randomly sampling inputs $\boldsymbol{x}$ from $p(\cdot)$. This third component enters the problem only through the combination of bias and random sampling, and so has not previously been considered in either the design or the active learning literature.

All of these components depend on $p(\boldsymbol{x})$, and also on the weights if weighted least squares (WLS) is carried out; the latter two components also depend on $\psi(\boldsymbol{x})$, which is necessarily unknown. We address this last point by maximizing the MSE over a full class of functions $\{\psi\}$, constrained only by certain identifiability and boundedness considerations; we then derive, by variational methods, input densities that minimize this maximized loss.

The boundedness of the effect of $\psi(\boldsymbol{x})$ must be treated carefully, in order that the components of the loss be of the same asymptotic order. We assume that

$$\int_{\chi} \psi^2(\boldsymbol{x})\, q(\boldsymbol{x})\, d\boldsymbol{x} \leq \tau_n^2 \tag{3}$$

for a given constant τ_n. Asymptotically, the measurement error results in contributions to the MSE of the order n^{-1}, through the variance of the parameter estimates. The squared bias is of the order τ_n^2 and the sampling (of $\boldsymbol{x}$) variation of the bias turns out to be of the order τ_n^2/n. At this point our development bifurcates. One approach, described in detail in Section 2 and solved in Section 3.1, is to employ weights—corresponding to importance weighting as discussed above, and also derived as "unbiased weights" in an experimental design context in Wiens (2000)—that asymptotically eliminate the bias, and to take $\tau_n^2 = O(1)$. We then optimize over the first and third components of the loss; in so doing we appear to be the first to address so explicitly the effect of sampling variation in active learning. A second possibility takes $\tau_n^2 = O\left(n^{-1}\right)$; in this case the sampling variation is asymptotically negligible in comparison with the first two components of the loss, over which we optimize in Section 3.2.

In expectations of the form $\int_\chi \cdots q(\boldsymbol{x})d\boldsymbol{x}$ (as in (3)) we use $d\boldsymbol{x}$ to represent either Lebesgue measure or counting measure. In the latter case $q(\boldsymbol{x})$ becomes a probability mass function and the integral becomes a weighted population average. When, as is invariably the case in practice, the underlying population, hence χ, is finite, we may take $q(\boldsymbol{x})$ to be the empirical mass function placing equal mass at each point in χ. Hence any expectations over $q(\cdot)$ become merely unweighted population averages.

Although we treat only approximately "linear" models in this article, the theory extends, with only minor modifications, to "nonlinear" models when one seeks "local optimality" in a neighbourhood of a linear approximation to the model. Kanamori (2002) and Sugiyama (2006) suggest taking a preliminary sample, in such cases, in order to obtain estimates to use as local parameters. An interesting alternate approach—see Krause et al. (2008)—is to (also) seek optimal robustness in the choice of the local parameter.

The individual problems addressed in this article are formulated in Section 2, with explicit solutions given in Sections 3.1 and 3.2. In Section 4 we discuss some of the computational aspects of the implementation of our methods, and carry out a Monte Carlo study.

In Section 5 we give the details of this approach in the context of benchmark data sets; in Section 6 in the context of a case study described in Section 1.1 below. We find that one can achieve significant reductions in the loss, relative to passive learning or to previously proposed methods of active learning. Our selection criteria all become functions of the "leverage values" commonly used as measures of influence in regression. More precisely if a sample is to be drawn from a population $\{\boldsymbol{x}_i | i = 1, \ldots, N\}$ and then the response queried, and if $\mathbf{H}$ is the "hat" matrix arising in a regression using the entire set $\{\boldsymbol{f}(\boldsymbol{x}_i) | i = 1, \ldots, N\}$ of possible regressors, then our methods are functions of $s(\boldsymbol{x}_i) = Nh_{ii}$.

We recommend three methods. One ("MVU") draws the sample with weights proportional to $\sqrt{s(\boldsymbol{x}_i)}$, and then downweights these points, using weights proportional to $1/\sqrt{s(\boldsymbol{x}_i)}$, in the regression. Another ("MVB") replaces $\sqrt{s(\boldsymbol{x}_i)}$ by $s(\boldsymbol{x}_i)$ itself; a third ("MMV") employs a more complicated function of $s(\boldsymbol{x}_i)$ requiring the computation of tuning constants.

The proofs of major mathematical results are in the Appendix. All computations were carried out in MATLAB; the code is available from us.

1.1. Case Study—Description

To motivate and illustrate the use of our methods, we shall in Section 6 consider a data set maintained by Statistics Canada (2015) and consisting of some 50,000 records containing ice thickness and snow depth measurements from a number of sites. Record length varies from station to station with some of the Arctic stations exceeding 50 years of observations. Measurements, rounded to the nearest centimeter, are taken at approximately the same location every year on a weekly basis, starting after freeze-up when the ice is safe to walk on, and continuing until break-up or when the ice becomes unsafe. The location is selected close to shore, but over a depth of water which will exceed the maximum ice thickness.

The importance of the relationship between the ice thickness and snow depth is discussed in Brown & Duguay (2011); we mention in particular that snow cover is related to the time of break-up, hence to the length of the shipping season. As measuring snow depth is relatively easy but measuring ice thickness generally requires augering holes in the ice, there is an obvious benefit in predicting the latter from the former. So in our study we seek to identify a subpopulation of snow records, sampling from which will yield results more precisely and more economically (fewer holes to be augered) than sampling the entire population.

2. PROBLEM FORMULATION

In (1), $\boldsymbol{f}(\boldsymbol{x})$ is a d-vector of functionally independent regressors, each element of which is a function of predictors $\boldsymbol{x} = (x_1, \ldots, x_k)'$, with $\boldsymbol{x}$ to be chosen from an "input space" χ, over which it has a density or mass function $q(\boldsymbol{x})$. Note that in the alternative models (2) the parameter $\boldsymbol{\theta}$ is not identifiable—an arbitrary linear combination of the regressors may be added to $\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\theta}$ and subtracted from $\psi(\boldsymbol{x})$; as a remedy we "define" this target parameter by

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\eta}} \int_{\chi} \left(E[Y|\boldsymbol{x}] - \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\eta}\right)^2 q(\boldsymbol{x})\, d\boldsymbol{x},$$

and then set $\psi(\boldsymbol{x}) = E[Y|\boldsymbol{x}] - \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\theta}$, as is dictated by (2). The minimization leads to the orthogonality requirement

$$\int_{\chi} \boldsymbol{f}(\boldsymbol{x})\,\psi(\boldsymbol{x})\,q(\boldsymbol{x})\,d\boldsymbol{x} = \boldsymbol{0}; \tag{4}$$

and we also assume (3).

Suppose now that $\boldsymbol{\theta}$ is estimated by WLS, with weights $\{w(\boldsymbol{x}_i)\}$. (We allow the possibility of ordinary least squares (OLS), in which case $w(\boldsymbol{x}_i) \equiv 1$.) A sample of size n is drawn from a density $p(\boldsymbol{x})$ on χ, and independent observations

$$Y_i = E[Y|\boldsymbol{x}_i] + \varepsilon_i, \quad i = 1, \ldots, n$$

are made; here $\{\varepsilon_i\}$ are i.i.d. random errors with mean zero and variance σ_ε^2.

Define matrices and vectors

$$\begin{aligned}
\boldsymbol{M}_{n;w} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{f}(\boldsymbol{x}_i)\,w(\boldsymbol{x}_i)\,\boldsymbol{f}'(\boldsymbol{x}_i) : d \times d,\\
\boldsymbol{D}_{n;w} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{f}(\boldsymbol{x}_i)\,w^2(\boldsymbol{x}_i)\,\boldsymbol{f}'(\boldsymbol{x}_i) : d \times d,\\
\boldsymbol{b}_{n;\psi,w} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{f}(\boldsymbol{x}_i)\,w(\boldsymbol{x}_i)\,\psi(\boldsymbol{x}_i) : d \times 1.
\end{aligned}$$

We will sample in such a way that $\boldsymbol{M}_{n;w}$ is non-singular. Then the WLS estimate is

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{\text{WLS}} &= \boldsymbol{M}_{n;w}^{-1}\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{f}(\boldsymbol{x}_i)\,w(\boldsymbol{x}_i)\,Y_i\\
&= \boldsymbol{\theta} + \boldsymbol{M}_{n;w}^{-1}\boldsymbol{b}_{n;\psi,w} + \boldsymbol{M}_{n;w}^{-1}\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{f}(\boldsymbol{x}_i)\,w(\boldsymbol{x}_i)\,\varepsilon_i,
\end{aligned}$$

with mean vector, covariance matrix, and mean squared error matrix

$$E_\varepsilon\left[\hat{\boldsymbol{\theta}}_{\text{WLS}}\right] = \boldsymbol{\theta} + \boldsymbol{M}_{n;w,p}^{-1}\boldsymbol{b}_{n;\psi,w},$$

$$\text{COV}_{\varepsilon}\left[\hat{\boldsymbol{\theta}}_{\text{WLS}}\right] = \frac{\sigma_{\varepsilon}^{2}}{n}\boldsymbol{M}_{n;w}^{-1}\boldsymbol{D}_{n;w}\boldsymbol{M}_{n;w}^{-1},$$

$$\text{MSE}_{\varepsilon}\left[\hat{\boldsymbol{\theta}}_{\text{WLS}}\right] = \boldsymbol{M}_{n;w}^{-1}\left\{\frac{\sigma_{\varepsilon}^{2}}{n}\boldsymbol{D}_{n;w} + \boldsymbol{b}_{n;\psi,w}\boldsymbol{b}_{n;\psi,w}'\right\}\boldsymbol{M}_{n;w}^{-1}.$$

Standard asymptotic theory—see Kanamori (2002) for some general asymptotic results in active learning—gives that

$$\boldsymbol{M}_{n;w} \overset{pr}{\to} \boldsymbol{M}_{w,p} = \int_{\chi} \boldsymbol{f}(\boldsymbol{x})\, w(\boldsymbol{x})\, \boldsymbol{f}'(\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x},$$

$$\boldsymbol{D}_{n;w} \overset{pr}{\to} \boldsymbol{D}_{w,p} = \int_{\chi} \boldsymbol{f}(\boldsymbol{x})\, w^{2}(\boldsymbol{x})\, \boldsymbol{f}'(\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x},$$

and, with

$$\boldsymbol{b}_{\psi,w,p} = \int_{\chi} \boldsymbol{f}(\boldsymbol{x})\, w(\boldsymbol{x})\, \psi(\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x},$$

that

$$\sqrt{n}\left(\boldsymbol{b}_{n;\psi,w} - \boldsymbol{b}_{\psi,w,p}\right) \overset{L}{\to} \boldsymbol{z} \sim N_{d}\left(\boldsymbol{0}, \boldsymbol{S}_{\psi,w,p}\right),$$

where the covariance matrix is

$$\begin{aligned}\boldsymbol{S}_{\psi,w,p} &= \text{COV}_{\boldsymbol{x}}\left[\boldsymbol{f}(\boldsymbol{x})\, w(\boldsymbol{x})\, \psi(\boldsymbol{x})\right] \\ &= \int_{\chi} \boldsymbol{f}(\boldsymbol{x})\, w^{2}(\boldsymbol{x})\, \psi^{2}(\boldsymbol{x})\, \boldsymbol{f}'(\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x} - \boldsymbol{b}_{\psi,w,p}\boldsymbol{b}_{\psi,w,p}'.\end{aligned}$$

Thus apart from terms that are $O_{p}\left(n^{-3/2}\right)$,

$$\text{MSE}_{\varepsilon}\left[\hat{\boldsymbol{\theta}}_{\text{WLS}}\right] \sim \frac{\sigma_{\varepsilon}^{2}}{n}\boldsymbol{M}_{w,p}^{-1}\boldsymbol{D}_{w,p}\boldsymbol{M}_{w,p}^{-1} + \boldsymbol{M}_{w,p}^{-1}\left(\boldsymbol{b}_{\psi,w,p} + \frac{1}{\sqrt{n}}\boldsymbol{z}\right)\left(\boldsymbol{b}_{\psi,w,p} + \frac{1}{\sqrt{n}}\boldsymbol{z}\right)'\boldsymbol{M}_{w,p}^{-1};$$

a further expectation over $\boldsymbol{z}$ results in

$$\text{MSE}_{\varepsilon,\boldsymbol{z}}\left[\hat{\boldsymbol{\theta}}_{\text{WLS}}\right] = \frac{1}{n}\boldsymbol{M}_{w,p}^{-1}\left(\sigma_{\varepsilon}^{2}\boldsymbol{D}_{w,p} + \boldsymbol{S}_{\psi,w,p}\right)\boldsymbol{M}_{w,p}^{-1} + \boldsymbol{M}_{w,p}^{-1}\boldsymbol{b}_{\psi,w,p}\boldsymbol{b}_{\psi,w,p}'\boldsymbol{M}_{w,p}^{-1} + O\left(n^{-3/2}\right).$$

A natural measure of the performance of the sampling density p is the integrated mean squared error (the "*full* expectation of the generalization error" in Sugiyama (2006, p. 142)), given by

$$\begin{aligned}\text{IMSE} &= \int_{\chi} E_{\varepsilon,\boldsymbol{z}}\left[\left\{\boldsymbol{f}'(\boldsymbol{x})\hat{\boldsymbol{\theta}} - E\left[Y|\boldsymbol{x}\right]\right\}^{2}\right] q(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \int_{\chi} \boldsymbol{f}'(\boldsymbol{x})\, \text{MSE}_{\varepsilon,\boldsymbol{z}}\left[\hat{\boldsymbol{\theta}}_{\text{WLS}}\right] \boldsymbol{f}(\boldsymbol{x})\, q(\boldsymbol{x})\, d\boldsymbol{x} + \int_{\chi} \psi^{2}(\boldsymbol{x})\, q(\boldsymbol{x})\, d\boldsymbol{x}.\end{aligned}$$

Here we again have used (4). With the definition

$$\boldsymbol{U} = \int_{\chi} \boldsymbol{f}(\boldsymbol{x})\, \boldsymbol{f}'(\boldsymbol{x})\, q(\boldsymbol{x})\, d\boldsymbol{x},$$

and ignoring terms that are $O\left(n^{-3/2}\right)$, this continues as

$$\text{IMSE} = tr\left(\boldsymbol{U}\text{MSE}_{\varepsilon,z}\left[\hat{\boldsymbol{\theta}}_{\text{WLS}}\right]\right) + \int_{\chi} \psi^2\left(\boldsymbol{x}\right) q\left(\boldsymbol{x}\right) d\boldsymbol{x}. \tag{5}$$

At this point we must specify the form of the dependence of τ_n^2 on n. As discussed in Section 1, a first option is to take constant $\tau_n \overset{def}{=} \tau$ and "unbiased" weights

$$w_0\left(\boldsymbol{x}\right) = \frac{q\left(\boldsymbol{x}\right)}{p\left(\boldsymbol{x}\right)}; \tag{6}$$

these yield $\boldsymbol{b}_{\psi,w_0,p} = \boldsymbol{0}$ by (4). (In all cases considered here, the weights are such that integrals in which they appear are finite.) As $\int_{\chi} \psi^2\left(\boldsymbol{x}\right) q\left(\boldsymbol{x}\right) d\boldsymbol{x}$ does not depend on the training density $p\left(\cdot\right)$ we drop it from (5) and work with n^{-1} times

$$\mathcal{L}_1(p|\psi) = n \cdot tr\left(\boldsymbol{U}\text{MSE}_{\varepsilon,z}\left[\hat{\boldsymbol{\theta}}_{\text{WLS}}\right]\right) = tr\left(\left[\sigma_\varepsilon^2 \boldsymbol{T}_p + \boldsymbol{S}_{\psi,p}\right]\boldsymbol{U}^{-1}\right). \tag{7}$$

To obtain (7) we have used that $\boldsymbol{M}_{w_0,p} = \boldsymbol{U}$ and employed the definitions

$$\boldsymbol{T}_p \overset{def}{=} \boldsymbol{D}_{w_0,p} = \int_{\chi} \boldsymbol{f}\left(\boldsymbol{x}\right)\boldsymbol{f}'\left(\boldsymbol{x}\right)\frac{q^2\left(\boldsymbol{x}\right)}{p\left(\boldsymbol{x}\right)} d\boldsymbol{x},$$

$$\boldsymbol{S}_{\psi,p} \overset{def}{=} \boldsymbol{S}_{\psi,w_0,p} = \int_{\chi} \boldsymbol{f}\left(\boldsymbol{x}\right)\frac{q^2\left(\boldsymbol{x}\right)}{p\left(\boldsymbol{x}\right)}\psi^2\left(\boldsymbol{x}\right)\boldsymbol{f}'\left(\boldsymbol{x}\right) d\boldsymbol{x}.$$

We note that this development corresponds to Equations (14)–(16) in Sugiyama (2006). The matrix $\boldsymbol{S}_{\psi,p}$ accounts for the variation in the conditional bias, which is a random variable due to the sampling from $p\left(\boldsymbol{x}\right)$. Sugiyama (2006) says that ψ is "inaccessible" and deals with it by taking $\tau_n = o(1)$, thus rendering the contribution of model misspecification, via $\boldsymbol{S}_{\psi,p}$, to the IMSE of a smaller order than random variation. Cohn, Ghahramani, & Jordan (1996, p. 131) also assume of their methods that "their squared bias is negligible when compared with their overall mean squared error," and go on to deal with random variation alone. Fukumizu (2000, p. 18) states that "we assume that the bias of the model is small enough to be neglected, and that active learning is supposed to reduce the variance term."

These methods of forcing or assuming both effects of model misspecification to be asymptotically negligible strike us as somewhat artificial, and we will instead adopt a minimax approach. In Section 3.1 we solve the problem:

P1) Find a sampling density p_1 for which $\sup_\psi \mathcal{L}_1(p_1|\psi) = \inf_p \sup_\psi \mathcal{L}_1(p|\psi)$, where the sup is evaluated subject to (4), and to (3) with $\tau_n = \tau$.

A second approach is to take $\tau_n = \tau/\sqrt{n}$, in which case $\boldsymbol{S}_{\psi,w,p} = O\left(n^{-1}\right)$ and, again ignoring terms that are $O\left(n^{-3/2}\right)$, (5) becomes n^{-1} times

$$\begin{aligned}\mathcal{L}_2(p|\psi, w) = {} & \sigma_\varepsilon^2 tr\boldsymbol{U}\boldsymbol{M}_{w,p}^{-1}\boldsymbol{D}_{w,p}\boldsymbol{M}_{w,p}^{-1} + \boldsymbol{b}'_{\sqrt{n}\psi,w,p}\boldsymbol{M}_{w,p}^{-1}\boldsymbol{U}\boldsymbol{M}_{w,p}^{-1}\boldsymbol{b}_{\sqrt{n}\psi,w,p} \\ & + \int_{\chi}\left(\sqrt{n}\psi\left(\boldsymbol{x}\right)\right)^2 q\left(\boldsymbol{x}\right) d\boldsymbol{x}.\end{aligned}$$

Note that although the effect of $\boldsymbol{S}_{\psi,w,p}$ has been eliminated, that of the bias has re-emerged. If unbiased weights (6) are used at this point then $\mathcal{L}_2(p|\psi, w_0) = \sigma_\varepsilon^2 \mathcal{L}_2(p) + \int_\chi \left(\sqrt{n}\psi(\boldsymbol{x})\right)^2 q(\boldsymbol{x})\, d\boldsymbol{x}$, where

$$\mathcal{L}_2(p) = tr\left(\boldsymbol{T}_p \boldsymbol{U}^{-1}\right) = \int_\chi \boldsymbol{f}'(\boldsymbol{x})\, \boldsymbol{U}^{-1} \boldsymbol{f}(\boldsymbol{x})\, \frac{q^2(\boldsymbol{x})}{p(\boldsymbol{x})} d\boldsymbol{x}.$$

Again we drop $\int_\chi \left(\sqrt{n}\psi(\boldsymbol{x})\right)^2 q(\boldsymbol{x})\, d\boldsymbol{x}$, which is bounded by τ^2 and does not depend on the training density. Note that $\mathcal{L}_2(q) = tr\boldsymbol{U}\boldsymbol{U}^{-1} = d$. In Section 3.2 we address the problem:

P2) Find a density p_2 for which $\mathcal{L}_2(p_2) = \min_p \mathcal{L}_2(p)$.

It will be seen that, although derived with respect to a different criterion, the solution to P2) coincides with that to P1) as $\sigma_\varepsilon^2/\tau^2 \to \infty$.

Remark 2.1. *A possible additional problem continues to take $\tau_n = \tau/\sqrt{n}$, but assumes the use of* OLS, *so that the effect of the bias remains non-negligible. This turns out to be very close to robust experimental design problems as reviewed in Wiens (2015). Some details are provided in Nie (2015); as the methods depend rather heavily on the structure of $q(\cdot)$ we do not pursue them here. A final problem of interest is to minimize the maximum mean squared error $\boldsymbol{f}'(\boldsymbol{t})\mathrm{MSE}_{\varepsilon,z}\left[\hat{\boldsymbol{\theta}}\right]\boldsymbol{f}(\boldsymbol{t})$ of fitted values $\hat{Y}(\boldsymbol{t}) = \boldsymbol{f}'(\boldsymbol{t})\hat{\boldsymbol{\theta}}$, with the maximum taken over $\boldsymbol{t} \in \chi$ as well as over ψ. See Nie (2015) for details in the straight line model $\boldsymbol{f}'(t) = (1, t)$.*

3. SOLUTIONS

We give the solutions to the problems formulated in the previous section; recall that these are characterized by the asymptotic order of the bound τ_n in (3).

3.1. Solution to P1): $\tau_n = O(1)$

In the Appendix we prove Lemma 1 and Theorem 1.

Lemma 1. *Recall (7). The supremum of $\mathcal{L}_1(p|\psi)$, over functions ψ satisfying (4), and (3) with $\tau_n = \tau$, is $\left(\sigma_\varepsilon^2 + \tau^2\right)$ times*

$$\mathcal{L}_1(p; \nu) = (1-\nu)\, E_q\left[a_p(\boldsymbol{x})\right] + \nu \sup_{\boldsymbol{x}} a_p(\boldsymbol{x}), \tag{8}$$

where $\nu = \tau^2/\left(\sigma_\varepsilon^2 + \tau^2\right) \in [0, 1]$ and $a_p(\boldsymbol{x}) \overset{def}{=} \frac{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}{p(\boldsymbol{x})} q(\boldsymbol{x})$.

Remark 3.1. *Note that ν may be chosen by the investigator, representing the relative concern for errors due to model misspecification rather than to variance of the response variable. Neither τ nor σ_ε need be to specified.*

Now $\mathcal{L}_1(p; \nu)$ is to be minimized over densities $p(\boldsymbol{x})$, for fixed ν. The solution is stated in the following theorem. The computational aspects are discussed in Section 4.

Theorem 1. *Define $s = \sup_\chi \boldsymbol{f}'(\boldsymbol{x})\, \boldsymbol{U}^{-1} \boldsymbol{f}(\boldsymbol{x})$ (which may be infinite) and*

$$m_0 = \sqrt{s} \int_\chi \sqrt{\boldsymbol{f}'(\boldsymbol{x})\, \boldsymbol{U}^{-1} \boldsymbol{f}(\boldsymbol{x})} q(\boldsymbol{x})\, d\boldsymbol{x}.$$

For $d \leq m < m_0$ define sets $\mathcal{A}_m$ of functions $a(\boldsymbol{x})$ by

$$\mathcal{A}_m = \left\{ a(\cdot) \mid \sup_{\boldsymbol{x}} a(\boldsymbol{x}) = m \text{ and } \int_{\chi} \frac{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}{a(\boldsymbol{x})} q(\boldsymbol{x})\, d\boldsymbol{x} = 1 \right\};$$

define as well

$$\mathcal{A}_{m_0} = \left\{ a(\cdot) \mid \sup_{\boldsymbol{x}} a(\boldsymbol{x}) \geq m_0 \text{ and } \int_{\chi} \frac{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}{a(\boldsymbol{x})} q(\boldsymbol{x})\, d\boldsymbol{x} = 1 \right\}.$$

For $m \in [d, m_0]$ define a member $a_m(\cdot)$ of $\mathcal{A}_m$ by

$$a_m(\boldsymbol{x}) = \min\left(c_m \sqrt{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}, m \right), \tag{9}$$

where c_m is defined by

$$\int_{\chi} \frac{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}{a_m(\boldsymbol{x})} q(\boldsymbol{x})\, d\boldsymbol{x} = 1. \tag{10}$$

Then $\sup_{\boldsymbol{x}} a_m(\boldsymbol{x}) = m$, and with the definition

$$m_\nu = \arg\min_{[d, m_0]} \left\{ (1-\nu) E_q[a_m(\boldsymbol{x})] + \nu m \right\}, \tag{11}$$

the density minimizing $\mathcal{L}_1(p; \nu)$ is

$$p_1(\boldsymbol{x}; \nu) = \frac{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}{a_{m_\nu}(\boldsymbol{x})} q(\boldsymbol{x}). \tag{12}$$

Minimax weights are (proportional to)

$$w(\boldsymbol{x}; \nu) = \frac{a_{m_\nu}(\boldsymbol{x})}{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}. \tag{13}$$

Remark 3.2. *When $\nu = 0$, $\mathcal{L}_1(p|\psi) = \mathcal{L}_2(p)$, whose minimizer is given in Section 3.2 and corresponds to $a_{m_0}(\boldsymbol{x})$ of Theorem 1. When $\nu = 1$, (11) gives $m_1 = d$, with $a_d(\boldsymbol{x}) \equiv d$ and $p_1(\boldsymbol{x}; 1) = \frac{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}{d} q(\boldsymbol{x})$. Simpler, direct proofs for both of these limiting cases are given in Nie (2015).*

Remark 3.3. *Note that the maximum loss for passive learning is $\mathcal{L}_1(q; \nu) = (1-\nu)d + \nu s$, which is typically infinite if $\nu > 0$ and χ is unbounded.*

3.2. Solution to P2): $\tau_n = O(1/\sqrt{n})$

This problem was essentially solved in Wiens (2000) in a different context; the solution can also be obtained as in Nie (2015): by the Cauchy–Schwarz inequality,

$$\mathcal{L}_2(p) = \int_{\chi} \left(\sqrt{\frac{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}{p(\boldsymbol{x})}} q(\boldsymbol{x}) \right)^2 d\boldsymbol{x} \cdot \int_{\chi} \left(\sqrt{p(\boldsymbol{x})} \right)^2 d\boldsymbol{x}$$
$$\geq \left(\int_{\chi} \sqrt{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})} q(\boldsymbol{x})\, d\boldsymbol{x} \right)^2 = \mathcal{L}_2(p_2).$$

The minimizing p_2 is

$$p_2(\boldsymbol{x}) = \frac{\sqrt{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}q(\boldsymbol{x})}{\int_{\chi}\sqrt{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}q(\boldsymbol{x})\,d\boldsymbol{x}}.$$

Note that

$$\mathcal{L}_2(p_2) < \int_{\chi}\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})q(\boldsymbol{x})\,d\boldsymbol{x} = d = \mathcal{L}_2(q).$$

4. COMPUTATIONAL ISSUES

In Sections 5 and 6 the population being considered is finite, of size N, say; as discussed in Section 1 any required expectations over $q(\cdot)$ then become unweighted population averages. In this context we will assess the following methods:

MVU ("Minimum Variance Unbiased") This is the solution to P2) and P1) with $\nu = 0$; we call it MVU as it minimizes the variance after using weights (6) to eliminate the mean conditional bias.

MMV ("Minimum Mixed Variance") This is the solution to P1) in which we take $\nu = 0.5$ so as to place equal weight on minimization of the estimation variance and of the variance of the conditional bias.

MVB ("Minimum Variance of the Bias") This is the solution to P1) with $\nu = 1$—all weight on minimization of the variance of the conditional bias.

For population values $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ define

$$\boldsymbol{U}_0 = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}'(\boldsymbol{x}_i), \text{ and } s(\boldsymbol{x}_i) = \boldsymbol{f}'(\boldsymbol{x}_i)\boldsymbol{U}_0^{-1}\boldsymbol{f}(\boldsymbol{x}_i).$$

With these definitions, and as pointed out in Section 1, $s(\boldsymbol{x}_i)/N = h_{ii}$, the ith leverage value in a regression using the entire population of predictors. From Section 3.2 and Remark 3.2, the ratios $r(\boldsymbol{x}) = p(\boldsymbol{x})/q(\boldsymbol{x})$ and weights $w(\boldsymbol{x}) = q(\boldsymbol{x})/p(\boldsymbol{x})$ for MVU and MVB are given by

MVU: $r_{MVU}(\boldsymbol{x}) = \sqrt{s(\boldsymbol{x})}\Big/\frac{1}{N}\sum_{i=1}^{N}\sqrt{s(\boldsymbol{x}_i)}$, $w_{MVU}(\boldsymbol{x}) = 1/r_{MVU}(\boldsymbol{x})$; and
MVB: $r_{MVB}(\boldsymbol{x}) = s(\boldsymbol{x})/d$, $w_{MVB}(\boldsymbol{x}) = 1/r_{MVB}(\boldsymbol{x})$.

Now for

MMV: To adapt the solution given in Section 3.1, first determine

$$m_\nu = \arg\min_{m\geq d}\left(m\left\{(1-\nu)\frac{1}{N}\sum_{i=1}^{N}\left[\min\left(t(m)\sqrt{s(\boldsymbol{x}_i)}, 1\right)\right] + \nu\right\}\right),$$

where $t(m)$ is defined by

$$1 = \frac{1}{N}\sum_{i=1}^{N}\frac{s(\boldsymbol{x}_i)}{m\cdot\min\left(t\sqrt{s(\boldsymbol{x}_i)}, 1\right)}.$$

Then with $\nu = 0.5$,

$$r_{MMV}(\boldsymbol{x}) = \frac{s(\boldsymbol{x})}{m_\nu \cdot \min\left(t(m_\nu)\sqrt{s(\boldsymbol{x})}, 1\right)}, w_{MMV}(\boldsymbol{x}) = \frac{1}{r_{MMV}(\boldsymbol{x})}.$$

Note that for each $\boldsymbol{x}$, $r_{MMV}(\boldsymbol{x})$ is either a multiple of $r_{MVU}(\boldsymbol{x})$ or a multiple of $r_{MVB}(\boldsymbol{x})$.

4.1. Monte Carlo

We have carried out a Monte Carlo study to assess empirically the properties of these procedures. In each of $L = 500$ simulations we generated "populations" of $N = 1{,}000$ i.i.d. input points X distributed as Student's t on 2 d.f. Then values of $Y = \theta_0 + \theta_1 x + \psi(x) + \varepsilon$ were obtained, with $\varepsilon \sim N(0, 1)$ and $\psi(x)$ quadratic, normed to satisfy (4) and $N^{-1}\sum \psi^2(x_i) = \tau = \nu/(1-\nu)$, for a range of values of ν. We then gathered samples of size $n = 100$, using probability weighted sampling and sampling ratios $r_{MVU}(\cdot)$, $r_{MMV}(\cdot)$, $r_{MVB}(\cdot)$, and $r_{Passive}(\cdot) \equiv 1$. From each of these samples we computed parameter estimates and predictions $\hat{Y} = \hat{\theta}_0 + \hat{\theta}_1 x$ for the unsampled population. We then computed the resulting estimates MSE: $\sum\left(\hat{Y}_i - Y_i\right)^2/(N-n)$ and prediction BIAS: $\sum\left(\hat{Y}_i - Y_i\right)/(N-n)$. In Table 1 we present the averages of these measures, over the L simulations, as well we give the standard errors of MSE.

All three of our active learning methods yielded substantially smaller mean squared errors than passive learning, with MMV and MVB being the best performers in this respect. Both were very stable over the range of values of ν. While having larger MSEs than our other two methods, MVU typically yielded a smaller prediction bias.

5. IMPLEMENTATIONS AND RECOMMENDATIONS; BENCHMARK DATA SETS

In this section, we apply our methods to eight benchmark data sets provided by DELVE (Rasmussen et al. 1996). They are labelled Bank-8fm, Bank-8fh, Bank-8nm, Bank-8nh, Kin-8fm, Kin-8fh, Kin-8nm, and Kin-8nh; each consists of $N = 8{,}192$ data points; and each data point consists of an 8-dimensional vector $\boldsymbol{x}$ of covariates and a univariate response Y, assumed hidden until queried. We first centred and standardized the predictors, so that the $N \times 8$ data matrix $\mathbf{X}$ has mean vector $N^{-1}\mathbf{X}'\mathbf{1}_N = \mathbf{0}$ and covariance matrix $(N-1)^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}_8$.

From each of the eight data sets we gather probability weighted training samples of size n, and fit a response $E[Y|\boldsymbol{x}] = \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\theta}$. The remaining responses are then predicted by $\hat{Y} = \boldsymbol{f}'(\boldsymbol{x})\hat{\boldsymbol{\theta}}$,

TABLE 1: Average (standard error) of MSE/prediction BIAS of learning methods.

	Method			
ν	MVU[a]	MMV[b]	MVB[c]	Passive
0.1	1.10 (0.011)/0.004	1.03 (0.002)/0.000	1.02 (0.002)/0.010	1.33 (0.043)/−0.018
0.3	1.27 (0.009)/0.006	1.04 (0.003)/0.010	1.05 (0.003)/0.019	2.24 (0.314)/−0.035
0.5	1.62 (0.022)/0.016	1.08 (0.004)/0.021	1.08 (0.004)/0.020	3.43 (0.256)/−0.040
0.7	2.47 (0.047)/0.019	1.16 (0.010)/0.022	1.16 (0.008)/0.033	6.12 (0.455)/−0.056
0.9	6.47 (0.204)/0.028	1.55 (0.032)/0.055	1.55 (0.030)/0.056	20.58 (2.48)/−0.144

[a] The solution to P2), and to P1) with $\nu = 0$.
[b] The solution to P1), with $\nu = 0.5$
[c] The solution to P1), with $\nu = 1$.

and the resulting MSE, as in Section 4.1, is computed. This process is carried out $L = 2{,}000$ times, allowing us to present the average of the L values of MSE, and the standard error of this average. We consider (i) $n = 20$—such small sample sizes might be appropriate if querying Y is very expensive—and a first-order model in which $\boldsymbol{f}(\boldsymbol{x})$ contains the $d = 9$ constant and linear regressors 1, $\left\{X_j | 1 \le j \le 8\right\}$; and (ii) $n = 200$ and a full second-order model ($d = 45$) that includes as well the quadratic terms $\left\{X_j X_k | 1 \le j \le k \le 8\right\}$.

The methods MVU, MMV, and MVB are compared with passive learning ($p = q$; $r(\boldsymbol{x}) \equiv 1$) and with "ALICE." The latter is a method proposed and preferred by Sugiyama (2006), after studying various methods of sampling for active learning, all assuming that $q(\boldsymbol{x})$ is Gaussian. When q is the $N(\mathbf{0}, \mathbf{I})$ density, p is taken to be a $N(\mathbf{0}, c^2\mathbf{I})$ density, so that

$$r(\boldsymbol{x}) \propto e^{\frac{1}{2}\left(1-\frac{1}{c^2}\right)\|\boldsymbol{x}\|^2}. \tag{14}$$

The value of c is chosen by comparing the choices $c = 0.70, 0.75, 0.80, \ldots, 2.4$ on preliminary samples. We found that if c is too far away from 1 then the weights (14) become very extreme, placing most weight on just a few points.

The results are reported in Tables 2 and 3. For each data set the best method, and any not significantly different from it (using individual two-sided t-tests with $\alpha = 0.05$) are shown in bold. (Some of the differences are so slight as to be of little practical significance but are declared significant because of the large value of L.) In the study recorded in Table 2 ($n = 20$, first order model), all three of our methods performed well on all data sets. ALICE was on a par with these methods in two cases, passive learning in none. In the study recorded in Table 3 ($n = 200$, second order model), the performance of MVU deteriorated slightly and that of ALICE improved slightly; that of the others was much as in the study of Table 2. The simplicity of MVU and MVB makes them recommended methods; although slightly harder to compute, MMV is an excellent performer that should certainly be in the toolbox as well.

TABLE 2: Average (standard error)[a] of MSEs of learning methods.
First-order model; $n = 20$.

	Method				
Data	MVU	MMV	MVB	ALICE[b]	Passive
Bank-8fm	**2.77 (0.021)**	**2.75 (0.021)**	**2.76 (0.020)**	3.16 (0.032)	2.80 (0.025)
Bank-8fh	**9.75 (0.084)**	**9.69 (0.076)**	**9.51 (0.075)**	10.24 (0.085)	10.05 (0.080)
Bank-8nm	**1.82 (0.017)**	**1.84 (0.016)**	**1.82 (0.0152)**	1.94 (0.019)	1.93 (0.018)
Bank-8nh	**6.31 (0.068)**	**6.23 (0.062)**	**6.14 (0.059)**	7.21 (0.080)	6.52 (0.061)
Kin-8fm	**0.84 (0.007)**	**0.83 (0.006)**	**0.84 (0.006)**	0.85 (0.007)	0.85 (0.007)
Kin-8fh	**3.73 (0.026)**	**3.68 (0.025)**	**3.67 (0.026)**	3.77 (0.026)	3.80 (0.0273)
Kin-8nm	**74.92 (0.557)**	**73.87 (0.512)**	**74.92 (0.529)**	**74.67 (0.505)**	75.79 (0.519)
Kin-8nh	**84.65 (0.619)**	**83.56 (0.585)**	**83.27 (0.598)**	**83.72 (0.584)**	85.02 (0.597)

[a] All values have been multiplied by 10^3.
[b] The preferred method of Sugiyama (2006).
For each data set the best method, and any not significantly different from it (using individual two-sided t-tests with $\alpha = 0.05$) are shown in bold.

TABLE 3: Average (standard error)[a] of MSEs of learning methods.
Second-order model; $n = 200$.

	Method				
Data	MVU	MMV	MVB	ALICE[b]	Passive
Bank-8fm	1.51 (0.003)	**1.45 (0.002)**	**1.45 (0.002)**	1.74 (0.006)	1.74 (0.006)
Bank-8fh	6.42 (0.010)	**6.36 (0.009)**	**6.36 (0.009)**	8.96 (0.035)	6.92 (0.015)
Bank-8nm	0.76 (0.002)	**0.70 (0.001)**	**0.70 (0.001)**	0.87 (0.003)	0.90 (0.003)
Bank-8nh	4.13 (0.007)	**4.04 (0.006)**	**4.05 (0.006)**	4.50 (0.011)	4.51 (0.011)
Kin-8fm	**0.191 (0.000)**	0.191 (0.000)	0.191 (0.000)	**0.191 (0.000)**	0.195 (0.000)
Kin-8fh	**2.26 (0.003)**	**2.26 (0.003)**	**2.27 (0.003)**	**2.25 (0.003)**	2.30 (0.003)
Kin-8nm	39.30 (0.055)	**38.94 (0.054)**	**39.05 (0.052)**	**38.88 (0.053)**	40.78 (0.066)
Kin-8nh	50.68 (0.067)	**50.58 (0.067)**	50.63 (0.065)	**50.45 (0.064)**	52.05 (0.074)

[a] All values have been multiplied by 10^3.

[b] The preferred method of Sugiyama (2006).

For each data set the best method, and any not significantly different from it (using individual two-sided t-tests with $\alpha = 0.05$) are shown in bold.

6. CASE STUDY—ANALYSIS

Brown & Duguay (2011, p. 2936) state that "Several studies have examined the effects on lake ice thickness through altering snow depths; finding overall, that decreasing the amount of snow cover resulted in thicker ice formation ...." Thus for the ice thickness data set described in Section 1.1 we transform to $X = (\text{snow depth} + 0.5)^{-1}$ and entertain a linear model relating ice thickness to X, and allowing for a different intercept and slope at each of the $g = 188$ sites for which there are complete records (N =48,597). Thus there are $d = 2g = 376$ parameters. The values of x were standardized, so as to have zero mean and unit variance.

To apply our methods we first computed $s\,(\boldsymbol{x}_i) = Nh_{ii}, i = 1, \ldots, N$; for this we took the qr-decomposition $\mathbf{F} = \mathbf{QR}$ of the $N \times d$ matrix $\mathbf{F}$, with rows $\boldsymbol{f}'(\boldsymbol{x}_i)$, and then computed the squared norms h_{ii} of the rows of $\mathbf{Q}$ (the entire hat matrix was much too large to be held in memory). For a range of sample sizes ("n") we then carried out weighted sampling from the population, with weights proportional to $r(\boldsymbol{x}_i)$, for each of the methods MVU, MMV, and MVB and for passive learning. For these samples the corresponding ice thicknesses were queried. This process was carried out 500 times, yielding the comparative values in Table 4.

In Figure 1 we have plotted the mean sampling weights in each site, against the number of locations per site; these range from 2 to 1,545. Perhaps not unexpectedly the smaller sites receive a disproportionate amount of the weight; this can be seen as justifying the stratification described below.

The following estimation scenarios were employed:

"ridge" Here we used least squares to compute the estimates from the samples; when the conditioning was so poor as to require it (as when $n < d$) we used ridge regression with tuning parameter $\lambda = 0.1$.

"lasso" When $n \ll d$ methods such as the lasso—see Hastie, Tibshirani, & Friedman (2009)—are commonly applied. The MATLAB implementation employed by us selects a regularization parameter λ adaptively.

TABLE 4: Average (standard error) of MSEs of learning methods used in the case study.

Using all sites ($N = 48{,}597$, $g = 188$)				
Unstratified sampling				
Method; sample size	MVU	MMV	MVB	Passive
Ridge; $n = 2{,}000$	0.54 (0.001)	0.53 (0.000)	0.53 (0.000)	0.71 (0.004)
Ridge; $n = 1{,}000$	0.75 (0.070)	0.80 (0.009)	0.85 (0.011)	0.96 (0.007)
Ridge; $n = 300$	1.15 (0.012)	1.28 (0.018)	1.32 (0.024)	1.09 (0.008)
Lasso; $n = 2{,}000$	8.88 (0.118)	1.54 (0.018)	1.71 (0.247)	2.55 (0.075)
Lasso; $n = 300$	64.50 (3.46)	46.15 (4.49)	24.28 (2.70)	55.35 (2.77)
Stratified sampling				
Ridge; $n = 2{,}000$	0.54 (0.001)	0.53 (0.001)	0.53 (0.000)	0.72 (0.004)
Ridge; $n = 1000$	0.62 (0.002)	0.59 (0.001)	0.59 (0.001)	0.93 (0.006)
Ridge; $n = 300$	1.00 (0.009)	1.02 (0.012)	0.89 (0.007)	1.14 (0.010)
Lasso; $n = 2{,}000$	28.50 (0.27)	15.50 (0.14)	17.67 (0.18)	2.21 (0.05)
Lasso; $n = 300$	81.13 (2.32)	56.93 (1.66)	45.27 (1.36)	20.67 (0.691)
Using only sites with at least 90 records ($N = 45{,}483$, $g = 100$)				
Unstratified sampling				
Method; sample size	MVU	MMV	MVB	Passive
Ridge; $n = 2{,}000$	0.63 (0.019)	0.57 (0.000)	0.59 (0.001)	1.45 (0.068)
Ridge; $n = 1{,}000$	1.58 (0.216)	0.83 (0.071)	0.74 (0.016)	0.97 (0.007)
Ridge; $n = 150$	1.15 (0.010)	1.17 (0.011)	1.19 (0.015)	1.13 (0.009)
Lasso; $n = 1{,}000$	17.44 (0.286)	3.60 (0.083)	3.36 (0.074)	5.11 (0.142)
Lasso; $n = 150$	114.79 (6.11)	61.91 (5.54)	48.78 (4.35)	70.63 (5.52)
Stratified sampling				
Ridge; $n = 2{,}000$	0.73 (0.038)	0.58 (0.006)	0.57 (0.003)	1.34 (0.025)
Ridge; $n = 1{,}000$	0.90 (0.101)	0.60 (0.001)	0.60 (0.001)	2.65 (0.577)
Ridge; $n = 150$	1.18 (0.010)	1.23 (0.014)	1.24 (0.018)	1.20 (0.012)
Lasso; $n = 1{,}000$	36.59 (0.46)	10.98 (0.14)	11.53 (0.19)	5.54 (0.17)
Lasso; $n = 150$	207.57 (9.08)	135.99 (6.62)	120.96 (6.19)	93.42 (4.09)

Both unstratified and stratified sampling—each probability weighted—were investigated. In the stratified approach the relative sample sizes in each site were as in the population, subject to rounding. The results in Table 4 confirm the advisability of the stratification—this almost always resulted in improved "ridge" estimates. Very typically, our weighting methods MVU, MMV, and MVB resulted in more efficient "ridge" predictions than passive learning. In regards to the use of the "lasso" our theory is not meant to cover this case and the large number of zeros among the parameter estimates led to very large MSEs when predicting the entire population, whether

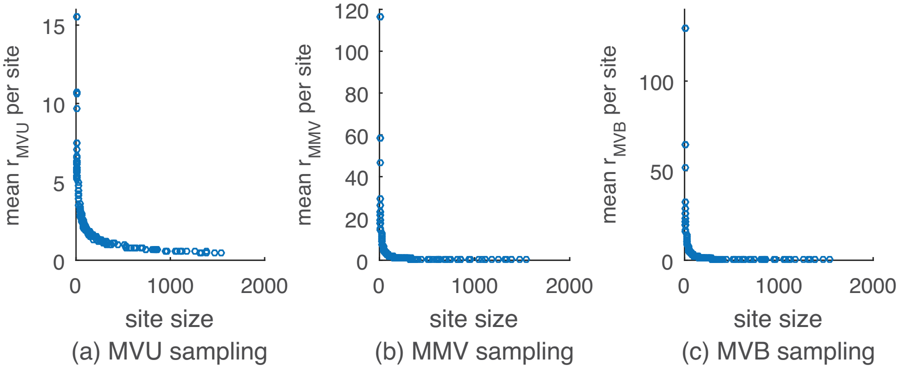

FIGURE 1: Case study: sampling weights r_{MVU}, r_{MMV}, r_{MVB} averaged in each site and plotted against site size.

active or passive learning was employed and both with and without stratification. Continuing to use ridge regression invariably led to much smaller MSEs.

APPENDIX

Proof of Lemma 1. Note that a maximizing ψ will necessarily attain equality in (3). Then $g(\boldsymbol{x}) = \psi^2(\boldsymbol{x})\, q(\boldsymbol{x}) / \tau^2$ is a density on χ, in terms of which

$$\begin{aligned}
\mathcal{L}_1(p|\psi) &= \left(\sigma_\varepsilon^2 + \tau^2\right) \cdot \left\{(1-\nu) E_q\left[a_p(\boldsymbol{x})\right] + \nu \int_\chi a_p(\boldsymbol{x}) \frac{\psi^2(\boldsymbol{x})\, q(\boldsymbol{x})}{\tau^2} d\boldsymbol{x}\right\} \\
&= \left(\sigma_\varepsilon^2 + \tau^2\right) \cdot \left\{(1-\nu) E_q\left[a_p(\boldsymbol{x})\right] + \nu \int_\chi a_p(\boldsymbol{x})\, g(\boldsymbol{x})\, d\boldsymbol{x}\right\} \\
&\le \left(\sigma_\varepsilon^2 + \tau^2\right) \cdot \left\{(1-\nu) E_q\left[a_p(\boldsymbol{x})\right] + \nu \sup_{\boldsymbol{x}} a_p(\boldsymbol{x})\right\}. \qquad \text{(A.1)}
\end{aligned}$$

The lemma asserts that the upper bound (A.1) is sharp. To prove this we will construct a sequence $\{\psi_\sigma(\boldsymbol{x})\}$ for which

$$\int_\chi a_p(\boldsymbol{x})\, \psi_\sigma^2(\boldsymbol{x})\, q(\boldsymbol{x})\, d\boldsymbol{x} \to \tau^2 \sup_{\boldsymbol{x}} a_p(\boldsymbol{x}) \text{ as } \sigma \to 0.$$

For this, we first construct a sequence $\{h_\sigma(\boldsymbol{x})\}$ of densities satisfying

$$\tau^2 \int_\chi a_p(\boldsymbol{x})\, h_\sigma(\boldsymbol{x})\, d\boldsymbol{x} \to \tau^2 \sup_{\boldsymbol{x}} a_p(\boldsymbol{x}) \text{ as } \sigma \to 0, \qquad \text{(A.2)}$$

and then verify that, for an associated sequence $\{\psi_\sigma(\boldsymbol{x})\}$, we have that

$$\int_\chi a_p(\boldsymbol{x}) \left[\psi_\sigma^2(\boldsymbol{x})\, q(\boldsymbol{x}) - \tau^2 h_\sigma(\boldsymbol{x})\right] d\boldsymbol{x} \to 0 \text{ as } \sigma \to 0. \qquad \text{(A.3)}$$

The construction of $\{h_\sigma(\boldsymbol{x})\}$ is detailed below. Given this sequence we set

$$\delta_\sigma(\boldsymbol{x}) = \sqrt{h_\sigma(\boldsymbol{x}) / q(\boldsymbol{x})},$$

$$\begin{aligned}
\boldsymbol{\alpha}_\sigma &= \int_\chi \boldsymbol{f}(\boldsymbol{x})\,\delta_\sigma(\boldsymbol{x})\,q(\boldsymbol{x})\,d\boldsymbol{x},\\
d_\sigma(\boldsymbol{x}) &= \boldsymbol{f}'(\boldsymbol{x})\,\boldsymbol{U}^{-1}\boldsymbol{\alpha}_\sigma,\\
\psi_\sigma(\boldsymbol{x}) &= \tau\frac{\delta_\sigma(\boldsymbol{x}) - d_\sigma(\boldsymbol{x})}{\sqrt{1-\boldsymbol{\alpha}'_\sigma\boldsymbol{U}^{-1}\boldsymbol{\alpha}_\sigma}}.
\end{aligned}$$

Then $\psi_\sigma(\boldsymbol{x})$, which is a projection of $\delta_\sigma(\cdot)$ onto the class of functions satisfying (4), satisfies (3) as well. When we construct $\{h_\sigma(\boldsymbol{x})\}$ we will also show that

$$\boldsymbol{\alpha}'_\sigma\boldsymbol{U}^{-1}\boldsymbol{\alpha}_\sigma \to 0 \text{ as } \sigma \to 0. \tag{A.4}$$

Condition (A.4) implies (A.3). To see this, assume for the moment that $\sup_{\boldsymbol{x}} a_p(\boldsymbol{x}) < \infty$. Then for σ sufficiently small that $\boldsymbol{\alpha}'_\sigma\boldsymbol{U}^{-1}\boldsymbol{\alpha}_\sigma < 1$ we have that

$$\begin{aligned}
&\left|\int_\chi a_p(\boldsymbol{x})\left[\psi^2_\sigma(\boldsymbol{x})\,q(\boldsymbol{x}) - \tau^2 h_\sigma(\boldsymbol{x})\right]d\boldsymbol{x}\right|\\
&= \left|\tau^2\int_\chi a_p(\boldsymbol{x})\left[\frac{(\delta_\sigma(\boldsymbol{x}) - d_\sigma(\boldsymbol{x}))^2}{1-\boldsymbol{\alpha}'_\sigma\boldsymbol{U}^{-1}\boldsymbol{\alpha}_\sigma} - \delta^2_\sigma(\boldsymbol{x})\right]q(\boldsymbol{x})\,d\boldsymbol{x}\right|\\
&\le \frac{\tau^2\sup_{\boldsymbol{x}} a_p(\boldsymbol{x})}{1-\boldsymbol{\alpha}'_\sigma\boldsymbol{U}^{-1}\boldsymbol{\alpha}_\sigma}\left\{\begin{array}{c}\boldsymbol{\alpha}'_\sigma\boldsymbol{U}^{-1}\boldsymbol{\alpha}_\sigma\int_\chi\delta^2_\sigma(\boldsymbol{x})\,q(\boldsymbol{x})\,d\boldsymbol{x} + \int_\chi d^2_\sigma(\boldsymbol{x})\,q(\boldsymbol{x})\,d\boldsymbol{x}\\ +2\int_\chi\delta_\sigma(\boldsymbol{x})\,|d_\sigma(\boldsymbol{x})|\,q(\boldsymbol{x})\,d\boldsymbol{x}\end{array}\right\}\\
&\le \frac{\tau^2\sup_{\boldsymbol{x}} a_p(\boldsymbol{x})}{1-\boldsymbol{\alpha}'_\sigma\boldsymbol{U}^{-1}\boldsymbol{\alpha}_\sigma}\left\{\begin{array}{c}\boldsymbol{\alpha}'_\sigma\boldsymbol{U}^{-1}\boldsymbol{\alpha}_\sigma\int_\chi\delta^2_\sigma(\boldsymbol{x})\,q(\boldsymbol{x})\,d\boldsymbol{x} + \int_\chi d^2_\sigma(\boldsymbol{x})\,q(\boldsymbol{x})\,d\boldsymbol{x}\\ +2\sqrt{\int_\chi\delta^2_\sigma(\boldsymbol{x})\,q(\boldsymbol{x})\,d\boldsymbol{x}}\sqrt{\int_\chi d^2_\sigma(\boldsymbol{x})\,q(\boldsymbol{x})\,d\boldsymbol{x}}\end{array}\right\}\\
&= \frac{\tau^2\sup_{\boldsymbol{x}} a_p(\boldsymbol{x})}{1-\boldsymbol{\alpha}'_\sigma\boldsymbol{U}^{-1}\boldsymbol{\alpha}_\sigma}\left\{\boldsymbol{\alpha}'_\sigma\boldsymbol{U}^{-1}\boldsymbol{\alpha}_\sigma + \boldsymbol{\alpha}'_\sigma\boldsymbol{U}^{-1}\boldsymbol{\alpha}_\sigma + 2\sqrt{\boldsymbol{\alpha}'_\sigma\boldsymbol{U}^{-1}\boldsymbol{\alpha}_\sigma}\right\},
\end{aligned}$$

and (A.3) follows. If $\sup_{\boldsymbol{x}} a_p(\boldsymbol{x}) = \infty$, replace $a_p(\boldsymbol{x})$ by $a_{p,m}(\boldsymbol{x}) = \min(a_p(\boldsymbol{x}), m)$ in the argument above, to conclude that $\int_\chi a_{p,m}(\boldsymbol{x})\psi^2_\sigma(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}$ and $\tau^2\int_\chi a_{p,m}(\boldsymbol{x})\,h_\sigma(\boldsymbol{x})\,d\boldsymbol{x}$ may be made arbitrarily close for each m, hence that both $\to\infty$ as $m\to\infty$.

To exhibit $\{h_\sigma(\boldsymbol{x})\}$ we assume that $\sup_{\boldsymbol{x}} a_p(\boldsymbol{x}) = a_p(\boldsymbol{x}^*)$ is attained. If it is not, then one may use instead a doubly indexed sequence $\left\{h_{\sigma,k}(\boldsymbol{x})\right\}$ with $\lim_k\lim_\sigma\int_\chi a_p(\boldsymbol{x})\,h_{\sigma,k}(\boldsymbol{x})\,d\boldsymbol{x} = \lim_k a_p(\boldsymbol{x}_k) = \sup_{\boldsymbol{x}} a_p(\boldsymbol{x})$.

To be specific, take

$$h_\sigma(\boldsymbol{x}) = c_\sigma\cdot\frac{1}{\sigma^k}\phi_k\left(\frac{\boldsymbol{x}-\boldsymbol{x}^*}{\sigma}\right) = c_\sigma\cdot\frac{1}{\sigma^k(2\pi)^{k/2}}e^{-\frac{\|\boldsymbol{x}-\boldsymbol{x}^*\|^2}{2\sigma^2}},$$

the k-dimensional normal density, with mean $\boldsymbol{x}^*$ and marginal variance σ^2 in each dimensional, with c_σ chosen so that the mass on χ is one. Then (A.2) is immediate. For (A.4) note that with $\beta = 2\sigma$ we have

$$\sqrt{h_\sigma(\boldsymbol{x})} = \sqrt{\sigma^k c_\sigma}2^k(2\pi)^{k/4}\cdot\frac{1}{\beta^k}\phi_k\left(\frac{\boldsymbol{x}-\boldsymbol{x}^*}{\beta}\right),$$

so that, with $E_\beta[\cdot]$ denoting expectation with respect to this k-dimensional normal density with mean $\boldsymbol{x}^*$ and marginal variance $\beta^2 \to 0$, we have that

$$\begin{aligned}\boldsymbol{\alpha}_\sigma &= \int_\chi \boldsymbol{f}(\boldsymbol{x})\,\delta_\sigma(\boldsymbol{x})\,q(\boldsymbol{x})\,d\boldsymbol{x}\\ &= \int_\chi \boldsymbol{f}(\boldsymbol{x})\sqrt{h_\sigma(\boldsymbol{x})\,q(\boldsymbol{x})}d\boldsymbol{x}\\ &= \sqrt{\sigma^k c_\sigma 2^k}\,(2\pi)^{k/4}\,E_\beta\left[\boldsymbol{f}(\boldsymbol{x})\sqrt{q(\boldsymbol{x})}\right]\\ &\to \mathbf{0},\end{aligned}$$

as $c_\sigma = O(1)$ and $E_\beta\left[\boldsymbol{f}(\boldsymbol{x})\sqrt{q(\boldsymbol{x})}\right] \to \boldsymbol{f}(\boldsymbol{x}^*)\sqrt{q(\boldsymbol{x}^*)}$ as $\sigma \to 0$. ■

Proof of Theorem 1. First note that if $m < d$ then the integral in the definition of $\mathcal{A}_m$ is $> d^{-1}\int_\chi \boldsymbol{f}'(\boldsymbol{x})\,\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})\,q(\boldsymbol{x})\,d\boldsymbol{x} = 1$, hence the restriction to $m \geq d$. Now define a set of densities by

$$\mathcal{P}_m = \left\{p(\cdot) \mid p(\boldsymbol{x}) = \frac{\boldsymbol{f}'(\boldsymbol{x})\,\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}{a(\boldsymbol{x})}q(\boldsymbol{x}) \mid a \in \mathcal{A}_m\right\}.$$

The set of all densities on χ coincides with $\cup_{m\in[d,m_0]}\mathcal{P}_m$ and so it suffices to minimize (8) over $\cup_{m\in[d,m_0]}\mathcal{A}_m$ and to then recover p, viz. to find

$$a_m = \arg\min_{a\in\mathcal{A}_m}\left\{(1-\nu)\,E_q[a(\boldsymbol{x})] + \nu\sup_{\boldsymbol{x}} a(\boldsymbol{x})\right\},$$

to then obtain m_ν from (11), and to recover $p_1(\boldsymbol{x};\nu)$ from (12).

We first verify that $a_m(\cdot) \in \mathcal{A}_m$ for $m \in [d, m_0]$. If $m < m_0$ then $\sup_{\boldsymbol{x}} a_m(\boldsymbol{x}) = m$ as long as $c_m \geq m/\sqrt{s}$, and then $\int_\chi \frac{\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}{a_m(\boldsymbol{x})}q(\boldsymbol{x})\,d\boldsymbol{x}$ ranges over $\left[d/m, m_0/m\right]$, so that there is a root c_m of (10). If $m = m_0$ then we set $c_{m_0} = \int_\chi \sqrt{\boldsymbol{f}'(\boldsymbol{x})\,\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}q(\boldsymbol{x})\,d\boldsymbol{x} = m_0/\sqrt{s}$, so that the function defined by (9) is $a_{m_0}(\boldsymbol{x}) = m_0\sqrt{\boldsymbol{f}'(\boldsymbol{x})\,\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})/s}$, for which $\sup_{\boldsymbol{x}} a_{m_0}(\boldsymbol{x}) = m_0$ and (10) holds.

The preceding argument also verifies that $\sup_{\boldsymbol{x}} a_m(\boldsymbol{x}) = m$, and so it remains only to show that (9) furnishes the minimizer in $\mathcal{A}_m$. For this, write $a_m(\boldsymbol{x}) = \min(a_-(\boldsymbol{x};m), m)$ for $a_-(\boldsymbol{x};m) = c_m\sqrt{\boldsymbol{f}'(\boldsymbol{x})\,\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}$. Let $a(\cdot)$ be any other member of $\mathcal{A}_m$ and consider

$$l(a) = \int_\chi\left[a(\boldsymbol{x})\,q(\boldsymbol{x}) + c_m^2\frac{\boldsymbol{f}'(\boldsymbol{x})\,\boldsymbol{U}^{-1}\boldsymbol{f}(\boldsymbol{x})}{a(\boldsymbol{x})}q(\boldsymbol{x})\right]d\boldsymbol{x} = \int_\chi q(\boldsymbol{x})\left[a(\boldsymbol{x}) + \frac{a_-^2(\boldsymbol{x};m)}{a(\boldsymbol{x})}\right]d\boldsymbol{x}.$$

As $a(\cdot) \in \mathcal{A}_m$ we have that $l(a) = E_q[a(\boldsymbol{x})] + c_m^2$, and so, if we can show that

$$l(a) \geq l(a_m) \text{ forall } a(\cdot) \in \mathcal{A}_m, \tag{A.5}$$

we will be done. Here we use that $\sup_{\boldsymbol{x}} a(\boldsymbol{x}) \geq \sup_{\boldsymbol{x}} a_m(\boldsymbol{x})$ for all $m \in [d, m_0]$.

To verify (A.5) let $\mathcal{R}_1$ and $\mathcal{R}_2$ be the sets where $a_m(\boldsymbol{x}) < m$ and $a_m(\boldsymbol{x}) = m$, respectively; then

$$l(a) = \int_{\mathcal{R}_1}\left[a(\boldsymbol{x}) + \frac{a_-^2(\boldsymbol{x};m)}{a(\boldsymbol{x})}\right]q(\boldsymbol{x})\,d\boldsymbol{x} + \int_{\mathcal{R}_2}\left[a(\boldsymbol{x}) + \frac{a_-^2(\boldsymbol{x};m)}{a(\boldsymbol{x})}\right]q(\boldsymbol{x})\,d\boldsymbol{x}$$

$$= l_1(a) + l_2(a), \text{ say.}$$

Then, as $a_m = a_-$ in the first integral and $= m$ in the second, we have that

$$\begin{aligned}
l(a) - l(a_m) &= [l_1(a) - l_1(a_m)] + [l_2(a) - l_2(a_m)] \\
&= \left[\int_{\mathcal{R}_1} \left\{ \left[a(\boldsymbol{x}) + \frac{a_-^2(\boldsymbol{x}; m)}{a(\boldsymbol{x})} \right] - \left[a_-(\boldsymbol{x}) + \frac{a_-^2(\boldsymbol{x}; m)}{a_-(\boldsymbol{x}; m)} \right] \right\} q(\boldsymbol{x})\, d\boldsymbol{x} \right] \\
&\quad + \int_{\mathcal{R}_2} \left\{ \left[a(\boldsymbol{x}) + \frac{a_-^2(\boldsymbol{x}; m)}{a(\boldsymbol{x})} \right] - \left[m + \frac{a_-^2(\boldsymbol{x}; m)}{m} \right] \right\} q(\boldsymbol{x})\, d\boldsymbol{x}.
\end{aligned}$$

Expanding these integrands and simplifying gives

$$\begin{aligned}
l(a) - l(a_m) = \int_{\mathcal{R}_1} &\frac{q(\boldsymbol{x})}{a(\boldsymbol{x})} [a_m(\boldsymbol{x}) - a(\boldsymbol{x})]^2\, d\boldsymbol{x} \\
&+ \int_{\mathcal{R}_2} \frac{q(\boldsymbol{x})}{a(\boldsymbol{x})} \left\{ \left[\frac{a_-^2(\boldsymbol{x})}{m} - a(\boldsymbol{x}) \right] [m - a(\boldsymbol{x})] \right\} d\boldsymbol{x}.
\end{aligned}$$

In the second integral $m - a(\boldsymbol{x}) \geq 0$ as $a(\cdot) \in \mathcal{A}_m$. As $a_m(\boldsymbol{x}) = m$ it must be the case that $a_-(\boldsymbol{x}) \geq m$ and so $\left[\frac{a_-^2(\boldsymbol{x})}{m} - a(\boldsymbol{x})\right] \geq [m - a(\boldsymbol{x})]$. Thus

$$\begin{aligned}
l(a) - l(a_m) &\geq \int_{\mathcal{R}_1} [a_m(\boldsymbol{x}) - a(\boldsymbol{x})]^2 \frac{q(\boldsymbol{x})}{a(\boldsymbol{x})} d\boldsymbol{x} + \int_{\mathcal{R}_2} [m - a(\boldsymbol{x})]^2 \frac{q(\boldsymbol{x})}{a(\boldsymbol{x})} d\boldsymbol{x} \\
&= \int_{\chi} [a_m(\boldsymbol{x}) - a(\boldsymbol{x})]^2 \frac{q(\boldsymbol{x})}{a(\boldsymbol{x})} d\boldsymbol{x} \\
&\geq 0,
\end{aligned}$$

with equality iff $a(\boldsymbol{x}) \equiv a_m(\boldsymbol{x})$. This establishes (A.5) and completes the proof. Identity (13) is immediate from (6). ■

ACKNOWLEDGEMENTS

This research is supported by the Natural Sciences and Engineering Research Council of Canada. We are grateful for the incisive comments of two anonymous referees.

BIBLIOGRAPHY

Box, G. E. P. & Draper, N. R. (1959). A basis for the selection of a response surface design. *Journal of the American Statistical Association*, 54, 622–654.

Brown, L. C., & Duguay, C. R. (2011). A comparison of simulated and measured lake ice thickness using a shallow water ice profiler. *Hydrological Processes*, 25, 2932–2941.

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4, 129–145.

Fukumizu, K. (2000). Statistical active learning in multilayer perceptrons. *IEEE Transactions on Neural Networks*, 11, 17–26.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*, Springer-Verlag, New York.

Huber, P. J. (1975). Robustness and designs. In *A Survey of Statistical Design and Linear Models*, Srivastava J. N., editor. North Holland, Amsterdam, 287–303.

Kanamori, T. (2002). Statistical asymptotic theory of active learning. *Annals of the Institute of Statistical Mathematics*, 54, 459–475.

Kanamori, T., & Shimodaira, H. (2003). Active learning algorithm using the maximum weighted log-likelihood estimator. *Journal of Statistical Planning and Inference*, 116, 149–162.

Kim, S.-J. & Boyd, S. (2008). A minimax theorem with applications to machine learning, signal processing, and finance. *SIAM Journal of Optimization*, 19, 1344–1367.

Krause, A., McMahan, H. B., Guestrin, C., & Gupta, A. (2008). Robust submodular observation selection. *Journal of Machine Learning Research*, 9, 2761–2801.

Lanckriet, G. R. G., El Ghaoui, L., Bhattacharyya, C., & Jordan, M. I. (2002). A robust minimax approach to classification. *Journal of Machine Learning Research*, 3, 555–582.

Liu, A. & Ziebart, B. D. (2014). Robust classification under sample selection bias, In *Advances in Neural Information Processing Systems*, January edition, Vol. 1, Neural information processing systems foundation, 37–45, 28th Annual Conference on Neural Information Processing Systems (NIPS) 2014, Montreal, Canada, 8–13 December.

Nie, R. (2015). *Robust Active Learning*, University of Alberta M.Sc. thesis, Department of Mathematical and Statistical Sciences.

Rasmussen, C. E., Neal, R. M., Hinton, G. E., van Camp, D., Revow, M., Ghahramani, Z., Kustra, R., & Tibshirani, R. (1996). *The Delve Manual*, url http://www.cs.toronto.edu/˜delve/data/datasets.html.

Scheffer, T., Decomain, C., & Wrobel, S. (2001). Active hidden Markov models for information extraction. In *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg, 2189, 309–318.

Settles, B. (2009). *Active Learning Literature Survey*, Computer Sciences Technical Report 1648, University of Wisconsin-Madison.

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

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

Statistics Canada (2015). *Ice Thickness Program Collection*. [Online], Available at: http://donnees.ec.gc.ca/data/ice/products/ice-thickness-program-collection/?lang=en.

Sugiyama, M., Krauledat, M., & Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross-validation. *Journal of Machine Learning Research*, 8, 985–1005.

Tong, S., & Koller, D. (2002). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2, 45–66.

Tur, G., Hakkani-Tür, D., & Schapire, R. E. (2005). Combining active and semisupervised learning for spoken language understanding. *Speech Communication*, 45, 171–186.

Wen, J., Yu, C.-N., & Greiner, R. (2014). Robust Learning Under Uncertain Test distributions: Relating covariate shift to model misspecification. In *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, 2014.

Wiens, D. P. (2000). Robust weights and designs for biased regression models: least squares and generalized M-estimation. *Journal of Statistical Planning and Inference*, 83, 395–412.

Wiens, D. P. (2015). Robustness of design. In *Handbook of Design and Analysis of Experiments*, Chapman & Hall/CRC.

Received 30 November 2016
Accepted 17 April 2017